UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES

PAR
JÉRÉMIE RAINVILLE

MODÉLISATION DE DONNÉES SPATIALES À L'AIDE DE CHAMPS ALÉATOIRES BASÉS SUR LA COPULE KHI-DEUX

DÉCEMBRE 2017

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

L'auteur de ce mémoire ou de cette thèse a autorisé l'Université du Québec à Trois-Rivières à diffuser, à des fins non lucratives, une copie de son mémoire ou de sa thèse.

Cette diffusion n'entraîne pas une renonciation de la part de l'auteur à ses droits de propriété intellectuelle, incluant le droit d'auteur, sur ce mémoire ou cette thèse. Notamment, la reproduction ou la publication de la totalité ou d'une partie importante de ce mémoire ou de cette thèse requiert son autorisation.

AVANT-PROPOS

Lorsqu'est venu le temps de choisir un domaine d'études supérieures, il a été facile pour moi d'opter pour les mathématiques. En effet, leur richesse ainsi que leur complexité en font un monde fascinant et rempli de possibilités. Le domaine des statistiques s'est ensuite présenté lorsque j'ai découvert leurs vastes champs d'étude ainsi que le nombre exorbitant d'applications concrètes leur étant reliées. Il fallait cependant préciser quelque peu la recherche et le choix s'est arrêté sur la modélisation de données spatiales à l'aide de la copule khi-deux. Par ses aspects à la fois fondamentaux et concrets, ce sujet s'est révélé aussi intéressant que motivant.

Je tiens à remercier le professeur Jean-François Quessy de m'avoir proposé ce sujet et d'avoir dirigé mes recherches. Sa disponibilité, sa grande connaissance ainsi que son expérience m'ont été précieux tout au long de mes études et lors de la rédaction de ce mémoire. Son soutien n'est pas étranger à l'aboutissement de cette recherche. J'aimerais aussi exprimer ma gratitude envers Marie-Hélène Toupin pour son aide et sa contribution au projet, notamment pour m'avoir initié au logiciel *Matlab*. Je remercie aussi Nadia Ghazzali et Louis Houde d'avoir accepté d'évaluer mon mémoire.

Je souhaite également exprimer ma reconnaissance à l'Institut des sciences mathématiques du Québec ainsi qu'au Conseil de Recherche en Sciences Naturelles et en Génie du Canada pour leur soutien financier qui m'a permis de me concentrer à temps plein sur mes recherches. Une mention particulière aux membres du Département de mathématiques et d'informatique de l'Université du Québec à Trois-Rivières pour leur encadrement et leur environnement où il fait bon d'étudier.

Finalement, j'offre une pensée toute spéciale à ma famille et mes amis pour leur support et leurs encouragements constants pour toute la durée de mes études.

Table des matières

Liste des tableaux

Table des figures

Introduction

Tant que les lois mathématiques se réfèrent à la réalité, elles ne sont pas certaines, et tant qu'elles sont certaines, elles ne se réfèrent pas à la réalité.
(Albert Einstein)

La théorie des copules est une théorie relativement nouvelle dans le monde des probabilités et statistique. Le concept de copule a été introduit à la fin des années 1950, en 1959 précisément, par un mathématicien américain du nom de Abe Sklar, professeur de mathématiques appliquées à l'Institut de technologies de l'Illinois. Il a toutefois fallu attendre plusieurs années pour que ce concept soit utilisé de façon beaucoup plus régulière en statistique. En effet, c'est dans les années 1970 que certains mathématiciens comme Kimeldorf, Sampson et Deheuvels ont décidé de faire utilisation des copules dans leurs travaux de recherche.

L'étude systématique des copules et le développement d'une théorie s'y intéressant débutent au milieu des années 1980 avec la contribution de plusieurs chercheurs québécois, notamment Christian Genest et Louis-Paul Rivard. À la fin des années 1990, de nombreux livres paraissent sur le sujet et la théorie prend de l'ampleur, notamment grâce au nombre grandissant de gens qui s'y intéressent. Ce soudain intérêt pour cette

théorie réside dans le fait de la découverte de son utilisation dans certains secteurs appliqués, particulièrement dans le domaine des finances ainsi que pour la modélisation spatiale. C'est cette dernière application qui nous intéressera pour ce document.

Il existe différents types de copules mais ce mémoire porte principalement sur la copule Khi-deux. Cette classe particulière a été proposée dans [1] par Andràs Bàrdossy dans le cadre de ses travaux sur les modèles géostatistiques utilisant les copules. Ce type particulier de copules lui permettait alors de mieux décrire la dépendance spatiale des paramètres de qualité sur les eaux souterraines sur lesquels il travaillait. L'utilisation de cette copule a continué après sa découverte, notamment dans les travaux [19] et [18] car elle répond à de nombreuses problématiques.

Dans ce travail, nous proposerons d'utiliser la famille des copules Khi-deux dans le cadre d'une situation de modélisation spatiale. Plus précisément, nous proposerons quatre types de modèles spatio-temporels basés sur cette famille, chacun incorporant des caractéristiques propres aux phénomènes à étudier. Avant d'y parvenir, les éléments nécessaires à la compréhension de ce processus seront explorés en détails.

Le Chapitre 1 introduit le lecteur au concept de copules, en s'attardant à leurs principales propriétés et en décrivant quelques familles de modèles. Le Chapitre 2 traite en détails de la famille des copules Khi-deux, en commençant par leur construction et en étudiant ensuite plusieurs de leurs propriétés les plus importantes. Le Chapitre 3 développe de nouveaux modèles spatio-temporels basés sur la famille de copules Khi-deux et propose des stratégies pour l'estimation de leurs paramètres ; les résultats d'études de simulations pour attester de la validité des stratégies d'estimation sont également présentés et commentés. Le Chapitre 4 est consacré à l'illustration des méthodologies proposées dans ce mémoire à de vraies données spatio-temporelles de températures et de précipitations recueillies à douze stations météorologiques autour de la Baie de San Francisco, en Californie. Les preuves se retrouvent à l'Annexe A et à l'Annexe B.

Chapitre 1

Une brève introduction aux copules

1.1 Notions préalables de statistique mathématique

Avant de se lancer dans le vif du sujet, à savoir la théorie des copules, il convient de se remémorer quelques notions fondamentales de statistique mathématique ; celles-ci faciliteront la compréhension de ce mémoire. On reverra en particulier les définitions de variables aléatoires et de lois de probabilités, ainsi que leurs extensions au cas de vecteurs aléatoires, autant dans le cas discret que dans le cas continu. Le lecteur estimant avoir une bonne compréhension de ces sujets peut ainsi passer directement à la section suivante portant sur la théorie des copules.

1.1.1 Variables aléatoires et lois de probabilité

Généralement, lorsqu'on souhaite effectuer de la modélisation de données à partir de modèles probabilistes, on travaille (implicitement ou explicitement) avec la notion de variables aléatoires. La définition formelle est donnée dans la suite.

Soit Ω, l'ensemble des valeurs possibles d'une expérience probabiliste. Sur cet ensemble, on définit une mesure de probabilité $\mathbb{P}$ sur les éléments d'une certaine σ-algèbre $\mathcal{S}$ (sous-ensemble non-vide et stable par complémentaire et union dénombrable) de Ω de sorte que le triplet $(\Omega, \mathcal{S}, \mathbb{P})$ forme un espace de probabilité. Une *variable aléatoire* est alors définie comme une fonction $X : \Omega \to \mathbb{R}$ et son comportement probabiliste est donné pour tout $A \subseteq \mathbb{R}$ par $\mathbb{P}_X(X \in A) = \mathbb{P}(X^{-1}(A))$.

On dit qu'une variable aléatoire X est discrète si l'ensemble de ses valeurs possibles, noté $\mathbb{X}$, est dénombrable. Dans ce cas, la probabilité associée à chaque résultat élémentaire $x \in \mathbb{X}$ est donnée par la fonction de masse $f(x) = \mathbb{P}_X(X = x)$. Inversement, une fonction arbitraire f est une fonction de masse sur un ensemble $\mathbb{X}$ si et seulement si $f(x) \geq 0$ pour tout $x \in \mathbb{X}$ et

$$\sum_{x \in \mathbb{X}} f(x) = 1.$$

Si l'ensemble $\mathbb{X}$ des valeurs possibles de X n'est pas dénombrable, typiquement un intervalle de $\mathbb{R}$, on dit que X est une variable aléatoire continue. Dans ce cas, la probabilité que X appartienne à un ensemble $A \subseteq \mathbb{X}$ est donnée par l'intégrale

$$\mathbb{P}_X(X \in A) = \int_A f(x)\,\mathrm{d}x,$$

où f est appellée la densité de probabilité. Inversement, une fonction arbitraire f est une densité de probabilité si et seulement si $f(x) \geq 0$ pour tout $x \in \mathbb{R}$ et

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1.$$

Une autre façon de caractériser le comportement d'une variable aléatoire se fait via sa fonction de répartition, c'est-à-dire que pour $A_x = (-\infty, x] \subseteq \mathbb{R}$, on considère

$$F_X(x) = \mathbb{P}_X(X \in A_x).$$

On note généralement $F_X(x) = \mathbb{P}_X(X \leq x)$ pour $x \in \mathbb{X}$. Ainsi,

$$F(x) = \begin{cases} \displaystyle\sum_{s \leq x} f(s) & \text{si } X \text{ est discrète;} \\ \\ \displaystyle\int_{-\infty}^{x} f(s)\,\mathrm{d}s & \text{si } X \text{ est continue.} \end{cases}$$

À noter que l'on déduit la densité f d'une variable aléatoire continue à partir de sa fonction de répartition F via

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x).$$

1.1.2 Quelques lois univariées importantes

On dit qu'une variable aléatoire continue X est distribuée selon la loi Normale de moyenne $\mu \in \mathbb{R}$ et de variance $\sigma^2 > 0$ si sa densité est donnée par

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

On note alors $X \sim \mathbb{N}(\mu, \sigma^2)$, où μ est la moyenne de X et σ^2, sa variance. La distribution Normale est une des plus importantes et une des plus utilisées. À noter que si $X \sim \mathbb{N}(\mu, \sigma^2)$, alors la variable aléatoire centrée réduite définie par $Z = (X - \mu)/\sigma$ est telle que $Z \sim \mathbb{N}(0, 1)$. Le calcul de probabilités concernant X peut donc toujours s'effectuer à partir de la loi $\mathbb{N}(0, 1)$. Spécifiquement, on a

$$\mathbb{P}_X(a \leq X \leq b) = \mathbb{P}_Z\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right).$$

Une autre distribution importante en statistique est la loi Khi-deux. Celle-ci est caractérisée à partir de variables aléatoires Normales centrées réduites indépendantes. De façon précise, soient $Z_1, \ldots, Z_\nu$, des variables aléatoires indépendantes telles que $Z_j \sim \mathbb{N}(0, 1)$, $j \in \{1, \ldots, \nu\}$. Alors la variable aléatoire $X = Z_1^2 + \cdots + Z_\nu^2$ suit une

loi Khi-deux à ν degrés de liberté ; on note alors $X \sim \chi^2_\nu$.

La loi de Student, très utile en inférence statistique, possède une représentation basée sur les lois Normale et Khi-deux. Spécifiquement, soient les variables indépendantes $Z \sim \mathbb{N}(0,1)$ et $Y \sim \chi^2_\nu$. On dit que la variable aléatoire

$$X = \frac{Z}{\sqrt{Y/\nu}}$$

suit une *loi de Student à ν degrés de liberté* ; on note $X \sim T_\nu$. Visuellement, la densité de Student s'apparente à celle de la loi Normale, avec pour différence que celle-ci varie en fonction du nombre ν de degrés de liberté. Lorsque ce nombre est supérieur ou égal à 30, la distribution de Student est très prés de la distribution de la loi Normale centrée réduite, ce qui la rend utile pour approximer une telle distribution. À noter que les fonctions de répartition ont volontairement été laissées de côté pour les distributions de Student et Khi-deux compte tenu de leur complexité et de leur faible pertinence.

1.1.3 Vecteurs aléatoires, lois jointes et corrélation

On verra que la théorie des copules intervient dans le cas où on s'intéresse simultanément à plusieurs variables aléatoires. Dans ce cas, on considère la notion de vecteurs aléatoires. Formellement, on dit simplement que $(X_1, \ldots, X_d)$ est un vecteur aléatoire de dimension d si $X_1, \ldots, X_d$ sont des variables aléatoires. Pour les rappels qui suivent, on se concentrera sur le cas bidimensionnel où on a une paire (X, Y) de variables aléatoires continues qui prennent leurs valeurs dans des ensembles $\mathbb{X}$ et $\mathbb{Y}$, respectivement. De la même façon que dans le cas d'une variable aléatoire continue, on peut définir une densité h sur l'ensemble produit $\mathbb{X} \times \mathbb{Y} \subseteq \mathbb{R}^2$ telle que

$$\mathbb{P}_{X,Y}(X \in A, Y \in B) = \int_A \int_B h(s,t)\,\mathrm{d}t\,\mathrm{d}s.$$

En posant $A_x = (-\infty, x]$ et $B_y = (-\infty, y]$, la fonction de répartition bivariée associée à la densité h est définie par

$$H_{X,Y}(x,y) = \mathbb{P}_{X,Y}(X \in A_x, Y \in B_y) = \int_{-\infty}^{x} \int_{-\infty}^{y} h(s,t)\,\mathrm{d}t\,\mathrm{d}s.$$

On peut montrer que toute fonction de répartition bivariée $H_{X,Y}$ dont les marges sont $F_X = \mathbb{P}_X(X \leq x)$ et $G_Y = \mathbb{P}_Y(Y \leq y)$ est telle que pour tout $(x,y) \in \mathbb{R}^2$,

$$H_{\text{inf}}(x,y) \leq H_{X,Y}(x,y) \leq H_{\text{sup}}(x,y), \tag{1.1}$$

où $H_{\text{inf}}(x,y) = \max\{F_X(x) + G_Y(y) - 1, 0\}$ et $H_{\text{sup}}(x,y) = \min\{F_X(x), G_Y(y)\}$. Pour démontrer ce résultat, considérons d'abord des événements A et B. On peut alors établir que $\mathbb{P}(A \cap B) \leq \min\{\mathbb{P}(A), \mathbb{P}(B)\}$, car à la fois $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ et $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$. De plus, les lois de *de Morgan* impliquent

$$\begin{aligned}
\mathbb{P}(A \cap B) &= 1 - \mathbb{P}(A^c \cup B^c) \\
&= 1 - \{\mathbb{P}(A^c) + \mathbb{P}(B^c) - \mathbb{P}(A^c \cap B^c)\} \\
&= \mathbb{P}(A) + \mathbb{P}(B) - 1 + \mathbb{P}(A^c \cap B^c) \\
&\geq \mathbb{P}(A) + \mathbb{P}(B) - 1,
\end{aligned}$$

où A^c et B^c représentent respectivement les complémentaires de A et de B. Comme une probabilité est nécessairement non-négative, on obtient $\mathbb{P}(A \cap B) \geq \max\{\mathbb{P}(A) + \mathbb{P}(B) - 1, 0\}$. Donc,

$$\max\{\mathbb{P}(A) + \mathbb{P}(B) - 1, 0\} \leq \mathbb{P}(A \cap B) \leq \min\{\mathbb{P}(A), \mathbb{P}(B)\}.$$

Ensuite, en posant $A_x = \{X \leq x\}$ et $B_y = \{Y \leq y\}$ et en remarquant que $\mathbb{P}(A_x) = F_X$ et $\mathbb{P}(B_y) = G_Y$, on tire que toute fonction de répartition bivariée H est telle que pour tout $(x,y) \in \mathbb{R}^2$,

$$\max\{F(x) + G(y) - 1, 0\} \leq H(x,y) \leq \min\{F(x), G(y)\}. \tag{1.2}$$

De manière similaire au cas univarié, la densité $h_{X,Y}$ de la paire (X, Y) peut s'obtenir en dérivant successivement $H_{X,Y}$ par rapport à x et à y, c'est-à-dire

$$h_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} H_{X,Y}(x, y).$$

Il est intéressant de remarquer que les comportements indivuels de X et de Y, appellés lois marginales, se déduisent de la loi jointe de (X, Y). Typiquement, on retrouve les fonctions de répartition marginales via les expressions

$$F_X(x) = \mathbb{P}_X(X \leq x) = \lim_{y \to \infty} H_{X,Y}(x, y) \quad \text{et} \quad F_Y(y) = \mathbb{P}_Y(Y \leq y) = \lim_{x \to \infty} H_{X,Y}(x, y).$$

Dans le cas où l'ensemble $\mathbb{Y}$ est borné supérieurement par b_Y, on a $F_X(x) = H_{X,Y}(x, b_Y)$. Similairement, si $\mathbb{X}$ est borné supérieurement par b_X, alors $F_Y(y) = H_{X,Y}(b_X, y)$. Une fois que les fonctions de répartition marginales sont obtenues, on déduit facilement les densités marginales via

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) \quad \text{et} \quad f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y).$$

Une fonction de répartition $H_{X,Y}$ permet de calculer toutes les probabilités relatives au comportement des variables de la paire (X, Y). En particulier, outre le fait que $\mathbb{P}_{X,Y}(X \leq x, Y \leq y) = H_{X,Y}(x, y)$, on peut montrer que pour tout $(x, y) \in \mathbb{X} \times \mathbb{Y}$,

$$\begin{aligned}
\mathbb{P}_{X,Y}(X \leq x, Y > y) &= F_X(x) - H_{X,Y}(x, y), \\
\mathbb{P}_{X,Y}(X > x, Y \leq y) &= F_Y(y) - H_{X,Y}(x, y), \\
\mathbb{P}_{X,Y}(X > x, Y > y) &= 1 - F_X(x) - F_Y(y) + H_{X,Y}(x, y).
\end{aligned}$$

La première formule est basée sur la constatation que $\{X \leq x\} = \{X \leq x \cap Y \leq y\} \cup \{X \leq x \cap Y > y\}$. En effet, comme les deux ensembles à droite de cette équation sont disjoints, on peut écrire

$$\mathbb{P}_X(X \leq x) = \mathbb{P}_{X,Y}(X \leq x, Y \leq y) + \mathbb{P}_{X,Y}(X \leq x, Y > y).$$

La deuxième identité s'obtient par des arguments similaires. La troisième relation est une conséquence d'une des lois de *de Morgan* qui stipule que toute mesure de probabilité $\mathbb{P}$ est telle que

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) = 1 - \{\mathbb{P}(A^c) + \mathbb{P}(B^c) - \mathbb{P}(A^c \cap B^c)\}.$$

Ainsi, on peut écrire

$$\begin{aligned} \mathbb{P}_{X,Y}(X > x, Y > y) &= \mathbb{P}_{X,Y}(X > x \cap Y > y) \\ &= 1 - \{\mathbb{P}_X(X \leq x) + \mathbb{P}_Y(Y \leq y) - \mathbb{P}_{X,Y}(X \leq x, Y \leq y)\} \\ &= 1 - F_X(x) - F_Y(y) + H_{X,Y}(x, y). \end{aligned}$$

En fait, la probabilité $\mathbb{P}_{X,Y}(X > x, Y > y)$ est reliée à la définition de fonction de survie. De façon explicite, la fonction de survie de (X, Y) est définie par

$$\bar{H}_{X,Y}(x, y) = \mathbb{P}_{X,Y}(X > x, Y > y).$$

De là, on déduit que les fonctions de survie de X et de Y sont respectivement

$$\begin{aligned} \bar{F}_X(x) &= \mathbb{P}_X(X > x) = \lim_{y \to -\infty} \bar{H}_{X,Y}(x, y), \\ \bar{F}_Y(y) &= \mathbb{P}_Y(Y > y) = \lim_{x \to -\infty} \bar{H}_{X,Y}(x, y). \end{aligned}$$

Un aspect important en statistique multidimensionnelle est de mesurer le lien qui existe entre des variables aléatoires. À cette fin, le coefficient de corrélation est souvent employé. Formellement, la corrélation entre les variables X et Y est définie par

$$\rho_{X,Y} = \mathrm{E}\left\{\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right\},$$

où $\mu_X = \mathrm{E}(X)$, $\sigma_X^2 = \mathrm{var}(X)$, $\mu_Y = \mathrm{E}(Y)$ et $\sigma_Y^2 = \mathrm{var}(Y)$. De façon équivalente,

$$\rho_{X,Y} = \frac{\mathrm{E}(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y}.$$

Si X et Y sont indépendantes, alors cela signifie que $\mathbb{P}_{X,Y}(X \leq x, Y \leq y) = \mathbb{P}_X(X \leq x)\,\mathbb{P}_Y(Y \leq y)$. Écrit autrement, on a la factorisation $H_{X,Y}(x,y) = F_X(x)\,F_Y(y)$. Une des conséquences directes est que $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$, ce qui entraîne que $\rho_{X,Y} = 0$. Supposons maintenant que X et Y sont linéairement reliées, c'est-à-dire que $Y = \beta_0 + \beta_1 X$, où $\beta_1 \neq 0$. Dans ce cas, on obtient aisément $\mu_Y = \beta_0 + \beta_1 \mu_X$, $\sigma_Y = |\beta_1|\,\sigma_X$ et $\mathrm{E}(XY) = \beta_0 \mu_X + \beta_1 \mathrm{E}(X^2) = \beta_0 \mu_X + \beta_1(\sigma_X^2 + \mu_X^2)$. Ainsi,

$$\rho_{X,Y} = \frac{\beta_0 \mu_X + \beta_1(\sigma_X^2 + \mu_X^2) - \mu_X\,(\beta_0 + \beta_1 \mu_X)}{\sigma_X |\beta_1|\,\sigma_X} = \frac{\beta_1\,\sigma_X^2}{|\beta_1|\,\sigma_X^2} = \frac{\beta_1}{|\beta_1|}.$$

Par conséquent, $\rho_{X,Y} = 1$ si $\beta_1 > 0$, c'est-à-dire lorsque X et Y sont en liaison linéaire positive parfaite. De même, $\rho_{X,Y} = -1$ si $\beta_1 < 0$, c'est-à-dire lorsque X et Y sont en liaison linéaire négative parfaite.

1.1.4 Quelques lois multidimensionnelles importantes

Cette sous-section décrit les extensions à d dimensions des lois Normale et Student. On verra qu'à l'instar du cas univarié, cette dernière distribution est construite à partir de la loi Normale. En premier lieu, on définit la matrice de variance-covariance $\Sigma \in \mathbb{R}^{d \times d}$ d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$ telle que $\Sigma_{jj'} = \mathrm{cov}(X_j, X_{j'})$. En particulier, la diagonale contient les variances de $X_1, \ldots, X_d$ car $\Sigma_{jj} = \mathrm{cov}(X_j, X_j) = \mathrm{var}(X_j)$. Par définition, la matrice de variance-covariance est symétrique puisque $\Sigma_{jj'} = \mathrm{cov}(X_j, X_{j'}) = \mathrm{cov}(X_{j'}, X_j) = \Sigma_{j'j}$. De plus, on peut montrer qu'elle est définie positive, c'est-à-dire que $\mathbf{v}\,\Sigma\,\mathbf{v}^\top > 0$ pour tout vecteur non nul $\mathbf{v} = (v_1, \ldots, v_d)$.

La loi Normale à d dimensions est définie à partir d'un vecteur $\mathbf{Z} = (Z_1, \ldots, Z_d)$ de variables indépendantes et de loi Normale centrée réduite. Spécifiquement, soient un vecteur de moyennes $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ ainsi qu'une matrice $\Sigma \in \mathbb{R}^{d \times d}$ symétrique et définie positive. On dit que le vecteur aléatoire $\mathbf{X} = \Sigma^{1/2}\,\mathbf{Z} + \mu$ est distribué selon une loi Normale à d dimensions de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance Σ ; on

note alors $\mathbf{X} \sim \mathbb{N}_d(\boldsymbol{\mu}, \Sigma)$. En partant de la représentation de $\mathbf{X}$ en fonction de $\mathbf{Z}$, on peut déduire que la densité de la loi $\mathbb{N}_d(\boldsymbol{\mu}, \Sigma)$ est donnée pour $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ par

$$\phi_{\boldsymbol{\mu},\Sigma}(\mathbf{x}) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{d/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}) \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})^\top \right\},$$

où $\mathbf{x}^T$ est le transposé de $\mathbf{x}$. Ceci amène la fonction de répartition écrite implicitement sous la forme

$$\Phi_{\boldsymbol{\mu},\Sigma}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} \phi_{\boldsymbol{\mu},\Sigma}(s_1, \ldots, s_d) \, \mathrm{d}s_d \cdots \mathrm{d}s_1. \tag{1.3}$$

À l'instar du cas univarié, la loi de Student d-dimensionnelle est construite à partir de la distribution Normale. Soit donc $\mathbf{X} = (X_1, \ldots, X_d)$, un vecteur aléatoire de loi Normale de moyennes nulles, c'est-à-dire $\boldsymbol{\mu} = (0, \ldots, 0)$, et de matrice de variance-covariance Σ. Soit également une variable Y de loi Khi-deux à ν degrés de liberté. Alors le vecteur

$$\mathbf{T} = \frac{\mathbf{X}}{\sqrt{Y/\nu}} = \left(\frac{X_1}{\sqrt{Y/\nu}}, \ldots, \frac{X_d}{\sqrt{Y/\nu}} \right)$$

est distribué selon une loi de Student d-dimensionnelle à ν degrés de liberté et de matrice de variance-covariance Σ; on note alors $\mathbf{T} \sim T_{d,\nu,\Sigma}$. Par construction, les marges de cette distribution sont toutes des lois de Student univariées à ν degrés de liberté. La densité de la loi de Student d-dimensionnelle est

$$h_{\Sigma,\nu}(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{d/2}} |\Sigma|^{-1/2} \left(1 + \frac{\mathbf{x}\Sigma^{-1}\mathbf{x}^\top}{\nu} \right)^{-(\nu+d)/2},$$

où $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t$ est la fonction Gamma. Sa fonction de répartition est

$$H_{\Sigma,\nu}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} h_{\nu,\Sigma}(s_1, \ldots, s_d) \, \mathrm{d}s_d \cdots \mathrm{d}s_1. \tag{1.4}$$

1.2 Fondements et propriétés des copules

Comme on le verra dans cette sous-section, le concept de copule est intimement relié aux fonctions de répartition multidimensionnelles. D'une certaine manière, il s'agit d'un objet mathématique qui capte toute la dépendance dans un vecteur aléatoire multidimensionnel. Le passage par cette théorie comporte deux avantages principaux quant à la modélisation multidimensionnelle, à savoir qu'elle permet

(i) d'isoler les liens d'interdépendance entre des variables aléatoires, indépendamment de leurs comportements individuels ; cet aspect est notamment très utile pour la définition de mesures de dépendance ;

(ii) la construction de modèles avec les choix désirés de lois marginales et de structure de dépendance ;

Le point de départ de toute la théorie des copules est le Théorème de Sklar. Cette section est consacrée, entre autres, à ce résultat important en analyse multidimensionnelle. Également, on verra un grand nombre de propriétés des copules ; on retrouve la plupart de celles-ci dans les ouvrages de [16] et [8].

1.2.1 Théorème de Sklar (1959)

Soit $H_{X,Y}$, une fonction de répartition bivariée de marges F_X et G_Y. Alors le Théorème de Sklar (voir [21]) stipule qu'il existe une fonction $C : [0,1]^2 \to [0,1]$ appelée la copule de $H_{X,Y}$ telle que pour tout $(x,y) \in \mathbb{R}^2$, on a

$$H_{X,Y}(x,y) = C\left\{F_X(x), G_Y(y)\right\}.$$

Si F_X et G_Y sont continues, alors la fonction C est unique et caractérise entièrement la dépendance entre des variables X et Y de loi $H_{X,Y}$.

On peut énoncer le Théorème de Sklar à partir de la fonction de survie $\bar{H}_{X,Y}(x,y) = \mathbb{P}_{X,Y}(X > x, Y > y)$. En effet, si $\bar{F}_X(x) = 1 - F_X(x)$ et $\bar{G}_Y = 1 - G_Y(y)$ sont les fonctions de survie univariées, alors on peut montrer qu'il existe une fonction $\widehat{C} : [0,1]^2 \to [0,1]$ telle que pour tout $(x,y) \in \mathbb{R}^2$, on a

$$\bar{H}_{X,Y}(x,y) = \widehat{C}\left\{\bar{F}(x), \bar{G}(y)\right\},$$

où $\widehat{C}(u,v) = u + v - 1 + C(1-u, 1-v)$ est la copule de survie de C. Pour montrer ce résultat, on note que puisque $H_{X,Y}(x,y) = C\{F_X(x), G_Y(y)\}$, on peut écrire

$$\begin{aligned}
\bar{H}_{X,Y}(x,y) &= 1 - F(x) - G(y) + H(x,y) \\
&= \{1 - F(x)\} + \{1 - G(y)\} - 1 + C\left\{F(x), G(y)\right\} \\
&= \bar{F}(x) + \bar{G}(y) - 1 + C\left\{1 - \bar{F}(x), 1 - \bar{G}(y)\right\} \\
&= \widehat{C}\left\{\bar{F}(x), \bar{G}(y)\right\}.
\end{aligned}$$

Le Théorème de Sklar permet de voir que la loi d'un couple de variables aléatoires X et Y peut s'exprimer en termes de ses fonctions de répartition marginales F_X et G_Y, et d'une copule C qui caractérise la dépendance entre X et Y. Ceci est très important quand vient le temps de faire de la modélisation, car on peut alors choisir des modèles univariés pour F_X et G_Y, indépendamment d'une structure de dépendance. On gagne ainsi énormément en flexibilité quant à la construction de modèles.

Il est intéressant de noter que la copule C d'un couple de variables aléatoires continues X et Y correspond à la loi conjointe de $U = F_X(X)$ et $V = G_Y(Y)$. En effet,

$$\begin{aligned}
\mathbb{P}\left\{F_X(X) \leq u, G_Y(Y) \leq v\right\} &= \mathbb{P}\left\{X \leq F_X^{-1}(u), Y \leq G_Y^{-1}(v)\right\} \\
&= H_{X,Y}\left\{F_X^{-1}(u), G_Y^{-1}(v)\right\} \\
&= C\left\{F_X \circ F_X^{-1}(u), G_Y \circ G_Y^{-1}(v)\right\} \\
&= C(u,v),
\end{aligned}$$

où $\circ$ est l'opérateur de composition de fonctions. On déduit de ce résultat que C est une fonction de répartition sur $[0,1]^2$ dont les marges sont uniformes. Cette dernière affirmation est simplement une conséquence du résultat bien connu à l'effet que les transformations intégrales de probabilité $F_X(X)$ et $G_Y(Y)$ sont uniformément distribuées sur $[0,1]$.

Il est possible d'étendre le Théorème de Sklar au cas d-dimensionnel. À cette fin, soit un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$ dont la fonction de répartition conjointe est $H_{\mathbf{X}}$ et dont les marges sont $F_{X_1}, \ldots, F_{X_d}$. Alors il existe une copule $C : [0,1]^d \to [0,1]$ telle que pour tout $(x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$H_{\mathbf{X}}(x_1, \ldots, x_d) = C\left\{F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right\}.$$

À l'instar du cas bidimensionnel, la copule C est unique lorsque toutes les lois marginales sont continues. On peut également établir que C correspond à la loi conjointe de $(U_1, \ldots, U_d)$, où $U_j = F_{X_j}(X_j)$ pour chaque $j \in \{1, \ldots, d\}$.

1.2.2 Invariance des copules

Une caractéristique importante concernant la copule de variables aléatoires est son invariance sous des transformations monotones croissantes. Formellement, on a dans le cas bidimensionnel que la copule C de (X, Y) est la même que la copule de $(\kappa(X), \eta(Y))$ lorsque κ et η sont des fonctions monotones croissantes. En effet,

$$\begin{aligned}
\mathbb{P}\left\{\kappa(X) \leq x, \eta(Y) \leq y\right\} &= \mathbb{P}\left\{X \leq \kappa^{-1}(x), Y \leq \eta^{-1}(y)\right\} \\
&= H_{X,Y}\left\{\kappa^{-1}(x), \eta^{-1}(y)\right\} \\
&= C\left\{F_X \circ \kappa^{-1}(x), G_Y \circ \eta^{-1}(y)\right\} \\
&= C\left\{\mathbb{P}\left(\kappa(X) \leq x\right), \mathbb{P}\left(\eta(Y) \leq y\right)\right\}.
\end{aligned}$$

Toutefois, si la transformation κ est plutôt monotone décroissante, alors la copule de $(\kappa(X), \eta(Y))$ est $\widetilde{C}(u,v) = v - C(1-u,v)$. En effet, on a dans ce cas que

$$\begin{aligned}
\mathbb{P}\{\kappa(X) \leq x, \eta(Y) \leq y\} &= \mathbb{P}\left\{X \geq \kappa^{-1}(x), Y \leq \eta^{-1}(y)\right\} \\
&= \mathbb{P}\left(Y \leq \eta^{-1}(y)\right) - \mathbb{P}\left\{X \leq \kappa^{-1}(x), Y \leq \eta^{-1}(y)\right\} \\
&= \mathbb{P}\left(Y \leq \eta^{-1}(y)\right) - C\left\{F_X \circ \kappa^{-1}(x), G_Y \circ \eta^{-1}(y)\right\} \\
&= \mathbb{P}\left(\eta(Y) \leq y\right) - C\left\{1 - \mathbb{P}\left(\kappa(X) \leq x\right), \mathbb{P}\left(\eta(Y) \leq y\right)\right\} \\
&= \widetilde{C}\left\{\mathbb{P}\left(\kappa(X) \leq x\right), \mathbb{P}\left(\eta(Y) \leq y\right)\right\}.
\end{aligned}$$

Par un raisonnement similaire, on trouve que la copule de $(\kappa(X), \eta(Y))$ quand κ est monotone croissante et η est monotone décroissante est $\widetilde{C}(u,v) = u - C(u, 1-v)$. Enfin, lorsque κ et η sont monotones décroissantes, on montre que la copule de $(\kappa(X), \eta(Y))$ est en fait la copule de survie $\widetilde{C}$ de C.

1.2.3 Extraction d'une copule

Il est possible d'extraire la copule associée à un couple (X, Y). Il s'agit en fait de poser $u = F_X(x)$ et $v = G_Y(y)$ dans l'équation $H_{X,Y}(x,y) = C\{F_X(x), G_Y(y)\}$, de telle sorte que $C(u,v) = H_{X,Y}\{F_X^{-1}(u), G_Y^{-1}(v)\}$. Dans le cas multidimensionnel, on a

$$C(u_1, \ldots, u_d) = H_{\mathbf{X}}\left\{F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(u_d)\right\}. \tag{1.5}$$

De façon générale, la densité d'une copule C, lorsqu'elle existe, est simplement

$$c(u_1, \ldots, u_d) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(u_1, \ldots, u_d).$$

À partir de l'Équation (1.5), on déduit l'expression

$$c(u_1, \ldots, u_d) = h_{\mathbf{X}}\left\{F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(u_d)\right\} \prod_{j=1}^{d} \frac{1}{f_{X_j} \circ F_{X_j}^{-1}(u_j)},$$

où $h_{\mathbf{X}}$ est la densité de $H_{\mathbf{X}}$ et $f_{X_1}, \ldots, f_{X_d}$ sont les densités marginales associées aux fonctions de répartitions $F_{X_1}, \ldots, F_{X_d}$.

1.2.4 Copule d'indépendance et bornes de Fréchet

Quand il n'existe aucun lien entre deux variables aléatoires X et Y, on dit qu'elles sont indépendantes. Dans ce cas,

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\,\mathbb{P}(Y \leq y) = \Pi\left\{\mathbb{P}(X \leq x), \mathbb{P}(Y \leq y)\right\},$$

où $\Pi(u, v) = uv$ est la copule d'indépendance. Dans le cas d-dimensionnel, la copule d'indépendance est $\Pi(u_1, \ldots, u_d) = u_1 \times \cdots \times u_d$.

Une dépendance positive parfaite entre deux variables aléatoires X et$Y$ survient lorsque $Y = \kappa(X)$ pour une certaine fonction κ qui est monotone croissante. Dans ce cas, on a $V = G_Y(Y) = F_X \circ \kappa^{-1}\{\kappa(X)\} = F_X(X) = U$. La copule de (X, Y) est donc

$$C(u, v) = \mathbb{P}(U \leq u, V \leq v) = \mathbb{P}(U \leq u, U \leq v) = \min(u, v).$$

Dans la suite, on notera $M(u, v) = \min(u, v)$ la copule de la dépendance positive parfaite. Lorsque que la relation fonctionnelle $Y = \kappa(X)$ survient pour une fonction κ qui est monotone décroissante, on parle de dépendance négative parfaite. On a alors $V = G_Y(Y) = 1 - F_X \circ \kappa^{-1}\{\kappa(X)\} = 1 - F_X(X) = 1 - U$. De là, on déduit que la copule de (X, Y) est

$$\begin{aligned} C(u, v) = \mathbb{P}(U \leq u, V \leq v) &= \mathbb{P}(U \leq u, 1 - U \leq v) \\ &= \mathbb{P}(U \leq u, U > -v + 1) \\ &= \mathbb{P}(U \leq u + v - 1) = \max(u + v - 1, 0). \end{aligned}$$

Dans la suite, on notera $W(u,v) = \max(u+v-1,0)$ la copule de la dépendance négative parfaite. Il est intéressant de noter que les inégalités de l'Équation (1.2) peuvent s'écrire sous la forme

$$W\left\{F_X(x), G_Y(y)\right\} \leq H_{X,Y}(x,y) \leq M\left\{F_X(x), G_Y(y)\right\}.$$

En utilisant le fait que $H_{X,Y}(x,y) = C\{F_X(x), G_Y(y)\}$ et en effectuant le changement de variables $u = F_X(x)$, $v = G_Y(y)$, on déduit que toute copule C satisfait les inégalités $W(u,v) \leq C(u,v) \leq M(u,v)$ pour tout $(u,v) \in [0,1]^2$. Les copules M et W se nomment respectivement les bornes inférieure et supérieure de Fréchet-Hoeffding.

1.2.5 Mesures de concordance

De façon traditionnelle, on utilise le coefficient de corrélation pour quantifier la force du lien entre deux variables aléatoires. Cependant, cette approche comporte son lot d'inconvénients. D'abord, il se peut que $\rho_{X,Y} = 0$ même si X et Y ne sont pas indépendantes. De plus, la valeur de $\rho_{X,Y}$ n'est pas invariante sous des transformations des marges. Autrement dit, il arrive que $\rho_{\kappa(X),\eta(Y)} \neq \rho_{X,Y}$, notamment lorsque κ et η sont des fonctions non-linéaires.

Pour combler ces lacunes, on aura recours à des mesures de dépendance qui s'écrivent en terme de la copule sous-jacente à une loi bidimensionnelle. Deux de ces mesures vont retenir notre attention, à savoir le tau de Kendall et le rho de Spearman. Celles-ci peuvent s'exprimer en fonction de l'opérateur de concordance Q entre deux distributions bivariées H et $H^\star$, à savoir

$$Q(H,H^\star) = \mathbb{P}\left\{(X-X^\star)(Y-Y^\star) > 0\right\} - \mathbb{P}\left\{(X-X^\star)(Y-Y^\star) < 0\right\},$$

où $(X,Y) \sim H$ et $(X^\star,Y^\star) \sim H^\star$ sont des paires indépendantes. En supposant que

H et $H^\star$ possèdent les mêmes marges F_X, G_Y et que celles-ci sont continues, alors

$$\begin{aligned} Q(H, H^\star) &= 2\,\mathbb{P}\left\{(X - X^\star)(Y - Y^\star) > 0\right\} - 1 \\ &= 2\,\mathbb{P}\left\{(F_X(X) - F_X(X^\star))\,(G_Y(Y) - G_Y(Y^\star)) > 0\right\} - 1 \\ &= 2\,\mathbb{P}\left\{(U - U^\star)(V - V^\star) > 0\right\} - 1, \end{aligned}$$

où $(U, V) = (F_X(X), G_Y(Y)) \sim C$ et $(U^\star, V^\star) = (F_X(X^\star), G_Y(Y^\star)) \sim C^\star$. On a donc $Q(H, H^\star) = Q(C, C^\star)$, où C et $C^\star$ sont les uniques copules associées aux lois H et $H^\star$, respectivement. Ainsi, l'opérateur de concordance entre deux lois bidimensionnelles est une notion qui réfère uniquement à leurs structures de dépendance sous-jacentes. À partir de la définition de concordance, on déduit par exemple de l'ouvrage de [16] que le tau de Kendall τ d'un couple (X, Y) de loi H et de copule C est défini par

$$\tau(C) = Q(C, C). \tag{1.6}$$

Pour sa part, le rho de Spearman ρ d'une copule C est donné par

$$\rho_S(C) = 3\,Q(C, \Pi). \tag{1.7}$$

1.2.6 Indices de dépendance codale

Pour une paire de variables aléatoires $(X, Y) \sim H_{X,Y}$ de fonctions marginales F_X et G_Y, les indices de dépendance codale inférieure et supérieure sont respectivement

$$\begin{aligned} \lambda_L &= \lim_{u \downarrow 0} \mathbb{P}\left\{X \leq F_X^{-1}(u) \middle| Y \leq G_Y^{-1}(u)\right\}, \\ \lambda_U &= \lim_{u \uparrow 1} \mathbb{P}\left\{X > F_X^{-1}(u) \middle| Y > G_Y^{-1}(u)\right\}. \end{aligned}$$

Si C est l'unique copule de $H_{X,Y}$ on a, en se rappelant que $(F_X(X), G_Y(Y)) \sim C$,

$$\begin{aligned}\lambda_L = \lambda_L(C) &= \lim_{u\downarrow 0} \frac{\mathbb{P}\left\{X \leq F_X^{-1}(u), Y \leq G_Y^{-1}(u)\right\}}{\mathbb{P}\left\{Y \leq G_Y^{-1}(u)\right\}} \\ &= \lim_{u\downarrow 0} \frac{\mathbb{P}\left\{F_X(X) \leq u, G_Y(Y) \leq u\right\}}{\mathbb{P}\left\{G(Y) \leq u\right\}} \\ &= \lim_{u\downarrow 0} \frac{C(u,u)}{u}.\end{aligned}$$

De manière similaire,

$$\begin{aligned}\lambda_U = \lambda_U(C) &= \lim_{u\uparrow 1} \frac{\mathbb{P}\left\{X > F_X^{-1}(u), Y > G_Y^{-1}(u)\right\}}{\mathbb{P}\left\{Y > G_Y^{-1}(u)\right\}} \\ &= \lim_{u\uparrow 1} \frac{\mathbb{P}\left\{F_X(X) > u, G_Y(Y) > u\right\}}{\mathbb{P}\left\{G_Y(Y) > u\right\}} \\ &= \lim_{u\uparrow 1} \frac{1 - 2u + C(u,u)}{1-u}.\end{aligned}$$

Ainsi, à l'instar du tau de Kendall et du rho de Spearman, les indices de dépendance codale sont des notions qui sont indépendantes des marges.

1.3 Quelques familles de copules

1.3.1 Copules Normales

La copule Normale est construite à partir de la fonction de répartition de la loi Normale standard définie à l'Équation (1.3). Pour la décrire, soit un vecteur aléatoire $\mathbf{X}$ de loi Normale d-dimensionnelle standard de matrice de variance-covariance Σ. Dans ce cas, comme $X_j \sim \mathbb{N}(0,1)$ pour chaque $j \in \{1,\ldots,d\}$, les lois marginales sont $\Phi(z) = \mathbb{P}(Z \leq z)$. En se basant sur la formule (1.5) pour l'extraction d'une copule,

on déduit que la copule Normale est définie par

$$C_{\Sigma}^{\mathrm{N}}(u_1,\ldots,u_d) = \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_d)} \phi_{\Sigma}(s_1,\ldots,s_d)\,\mathrm{d}s_d \cdots \mathrm{d}s_1,$$

où ϕ_{Σ} est la densité de la loi Normale à d dimensions. À partir de cette définition, on peut observer qu'il est facile à l'aide des copules de construire un modèle dans lequel la dépendance entre les différentes variables aléatoires est de type Normale et les fonctions marginales sont choisies de façon arbitraire. En effet, il s'agit de choisir des marges $F_{X_1},\ldots,F_{X_d}$ et de poser

$$H_{\Sigma,F_1,\ldots,F_d}(\mathbf{x}) = C_{\Sigma}^{\mathrm{N}}\left\{F_{X_1}(x_1),\ldots,F_{X_d}(x_d)\right\}.$$

Soient (X,Y) et (X^*,Y^*), des couples de variables aléatoires ayant respectivement pour copule C^{N} de corrélation ρ et C^{N} de corrélation $\widetilde{\rho}$. Il a été démontré par [14] que l'opérateur de concordance pour la copule Normale est donné par

$$Q\left(C_{\rho}^{\mathrm{N}}, C_{\widetilde{\rho}}^{\mathrm{N}}\right) = 4\,\Phi_{\rho^+}(0,0) - 1 = \frac{2}{\pi}\sin^{-1}\rho^+,$$

où $\rho^+ = (\rho+\widetilde{\rho})/2$. De là, les Équations (1.6)–(1.7) permettent d'obtenir

$$\tau(C_{\rho}^{\mathrm{N}}) = \frac{2}{\pi}\sin^{-1}\rho \quad \text{et} \quad \rho_S(C_{\rho}^{\mathrm{N}}) = \frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right).$$

Les indices de dépendance codale sont nuls pour la copule Normale lorsque $\rho < 1$, c'est-à-dire que $\lambda_L(C_{\rho}^{\mathrm{N}}) = \lambda_U(C_{\rho}^{\mathrm{N}}) = 0$. Pour $\rho = 1$, $\lambda_L(C_{\rho}^{\mathrm{N}}) = \lambda_U(C_{\rho}^{\mathrm{N}}) = 1$.

1.3.2 Copules de Student

En procédant de manière analogue à ce qui a été employé pour définir la copule Normale, on peut déduire la copule associée à loi de Student d-dimensionnelle à ν degrés de liberté. En effet, en se référant à l'Équation (1.4) et au fait que les fonctions

de répartition marginales sont toutes F_ν, à savoir la fonction de répartition de la loi de Student univariée à ν degrés de liberté, la copule de Student est donnée par

$$C_{\Sigma,\nu}(u_1,\ldots,u_d) = \int_{-\infty}^{F_\nu^{-1}(u_1)} \cdots \int_{-\infty}^{F_\nu^{-1}(u_d)} h_{\Sigma,\nu}(s_1,\ldots,s_d)\,\mathrm{d}s_d \cdots \mathrm{d}s_1,$$

où $h_{\Sigma,\nu}$ est la densité Student d-dimensionnelle à ν degrés de liberté et de matrice Σ.

1.3.3 Copules Archimédiennes

La classe des copules Archimédiennes est très utilisée en pratique car elle permet de modéliser la dépendance dans des modèles comportant un nombre élevé de dimensions de façon relativement aisée. Ces copules sont construites à partir d'un générateur Archimédien qiu respecte certaines propriétés. Plus précisément, un *générateur Archimédien* est une fonction $\phi : [0,1] \to [0,\infty]$ qui est continue, strictement décroisante, convexe et telle que $\phi(1) = 0$. De plus, ϕ^{-1} doit être une fonction d-monotone au sens où pour tout $j \in \{1,\ldots,d\}$,

$$(-1)^j \frac{\mathrm{d}^j}{\mathrm{d}t^j}\phi^{-1}(t) > 0.$$

Pour une fonction ϕ qui satisfait toute ces conditions, on définit alors qu'une copule Archimédienne est de la forme

$$C_\phi(u_1,\ldots,u_d) = \phi^{-1}\left\{\sum_{j=1}^{d}\phi(u_j)\right\}.$$

Par construction, les copules Archimédiennes sont échangeables car pour toute permutation $(\pi_1, \ldots, \pi_d)$ de l'ensemble $\{1, \ldots, d\}$ des d premiers entiers,

$$\begin{aligned} C_\phi\left(u_{\pi_1}, \ldots, u_{\pi_d}\right) &= \phi^{-1}\left\{\sum_{j=1}^{d} \phi\left(u_{\pi_j}\right)\right\} \\ &= \phi^{-1}\left\{\sum_{j=1}^{d} \phi(u_j)\right\} \\ &= C_\phi(u_1, \ldots, u_d). \end{aligned}$$

Dans le cas bidimensionnel, on a $C_\phi(u_1, u_2) = \phi^{-1}\{\phi(u_1) + \phi(u_2)\}$. Ces copules sont associatives au sens où pour tout $u_1, u_2, u_3 \in [0, 1]$,

$$C_\phi\{u_1, C_\phi(u_2, u_3)\} = C_\phi\{C_\phi(u_1, u_2), u_3\}.$$

On peut montrer que ϕ et $\widetilde{\phi} = K\,\phi$, où $K > 0$, génèrent la même copule Archimédienne. En effet, en notant que $\widetilde{\phi}^{-1}(u) = \phi^{-1}(u/K)$, on obtient

$$\begin{aligned} C_{\widetilde{\phi}}(u_1, \ldots, u_d) &= \widetilde{\phi}^{-1}\left\{\sum_{j=1}^{d} \widetilde{\phi}(u_j)\right\} \\ &= \widetilde{\phi}^{-1}\left\{K \sum_{j=1}^{d} \phi(u_j)\right\} \\ &= \phi^{-1}\left\{K \sum_{j=1}^{d} \phi(u_j) \Big/ K\right\} \\ &= \phi^{-1}\left\{\sum_{j=1}^{d} \phi(u_j)\right\} \\ &= C_\phi(u_1, \ldots, u_d). \end{aligned}$$

Avant de présenter quelques modèles paramétriques de copules Archimédiennes, on note que la copule d'indépendance fait partie de la famille Archimédienne. En effet, en posant $\phi(t) = -\ln t$, on déduit que $C_\phi(u_1, \ldots, u_d) = u_1 \times \cdots \times u_d$.

Exemple 1.1. *La copule de Clayton, issue d'un modèle étudié par [3], est générée*

par $\phi_\theta(t) = (t^{-\theta} - 1)/\theta$. *Quand* $\theta > 0$, *la forme de cette copule est*

$$C_\theta^{\mathrm{CL}}(u_1, \ldots, u_d) = \left(\sum_{j=1}^{d} u_j^{-\theta} - d + 1 \right)^{-1/\theta}.$$

On retrouve la copule d'indépendance à la limite quand $\theta \to 0$.

Exemple 1.2. *Une autre copule intéressante est celle de Gumbel–Hougaard générée pour* $\theta \in [0, 1]$ *par* $\phi_\theta(t) = (-\ln t)^{1/(1-\theta)}$. *Dans le cas bivariée, cette copule Archimédienne s'exprime par*

$$C_\theta(u_1, u_2) = \exp\left\{ -\left((-\ln u_1)^{1/(1-\theta)} + (-\ln u_2)^{1/(1-\theta)} \right)^{1-\theta} \right\}.$$

À l'instar de la copule de Clayton, on retrouve l'indépendance comme cas particulier lorsque $\theta = 0$. *À noter enfin que la copule de Gumbel–Hougaard fait également partie de la classe générale des copules à valeurs extrêmes.*

Il est possible de déterminer le générateur d'une copule C, sachant que celle-ci appartient à la famille des copules Archimédiennes. En effet, en supposant que son générateur est ϕ, on note que les dérivées partielles de C sont

$$\frac{\partial}{\partial u} C(u, v) = \frac{\phi'(u)}{\phi' \circ \phi^{-1} \{\phi(u) + \phi(v)\}} \quad \text{et} \quad \frac{\partial}{\partial v} C(u, v) = \frac{\phi'(v)}{\phi' \circ \phi^{-1} \{\phi(u) + \phi(v)\}},$$

de telle sorte que leur quotient est

$$\frac{\partial C(u, v)/\partial u}{\partial C(u, v)/\partial v} = \frac{\phi'(u)}{\phi'(v)}.$$

Ce quotient permet alors de déduire une forme pour ϕ', à une constante près. Il s'agit ensuite de résoudre une équation différentielle d'ordre un pour déduire le générateur ϕ. Une illustration de ce procédé est donnée dans l'exemple qui suit.

Exemple 1.3. *Pour la copule d'indépendance* $\Pi(u, v) = uv$, *on a*

$$\frac{\partial}{\partial u}\Pi(u, v) = v \quad \textit{et} \quad \frac{\partial}{\partial v}\Pi(u, v) = u.$$

Le générateur ϕ *de la copule* Π *est donc tel que* $\phi'(u)/\phi'(v) = v/u$. *On peut alors déduire que* $\phi'(u) = K_1/u$ *pour une certaine constante* K_1. *De là,*

$$\phi(u) = \int \phi'(u)\,\mathrm{d}u = \int \frac{K_1}{u}\,\mathrm{d}u = K_1 \ln u + K_2.$$

Puisque ϕ *doit satisfaire* $\phi(1) = 0$, *on a nécessairement* $K_2 = 0$. *De plus, comme* ϕ *doit être décroissante, il faut absolument que* $K_1 < 0$. *Enfin, comme les générateurs Archimédiens sont équivalents à une constante près, on peut prendre* $K_1 = -1$ *et conclure que* $\phi(t) = -\ln t$.

Chapitre 2

La famille générale des copules Khi-deux multidimensionnelles

2.1 Mise en contexte

Une famille de copules qui mérite une attention approfondie est la famille Khi-deux. En effet, ce sont ces copules qui nous intéresseront lorsque viendra le temps de présenter nos modèles pour modéliser des données spatiales. C'est pourquoi il est primordial de bien définir cette famille de copules et de décrire certaines de ses propriétés. Comme il a été mentionné précédemment, cette classe de copules a été décrite initialement par [1], puis utilisée par [10] et [18] pour modéliser des données spatiales. Nous verrons que cette copule représente une alternative intéressante à la copule Normale.

La copule Normale présente plusieurs points forts, ce qui la rend attirante dans plusieurs situations. En particulier, elle permet la modélisation en grandes dimensions et chaque paire est spécifiquement paramétrisée. De plus, le fait qu'elle utilise une matrice de corrélation la rend appropriée pour la statistique spatiale, car on peut

relier le niveau de dépendance entre deux stations en fonction de la distance qui les sépare. Cependant, la classe des copules Normales comporte aussi quelques désavantages. D'abord, elle ne permet pas de modéliser des structures de dépendance dont les queues inférieures et supérieures sont différentes ; en effet, la copule Normale possède la propriété de symétrie radiale, ce qui peut être assez limitatif en pratique. On verra dans ce chapitre que la famille des copules Khi-deux préserve les propriétés souhaitables de la copule Normale, tout en permettant de l'asymétrie radiale.

Dans ce chapitre, nous verrons donc les étapes pour construire la copule Khi-deux ; les cas bivarié et multidimensionnel seront traités séparément. Plusieurs propriétés de cette famille de modèles seront décrites, incluant le calcul de mesures de dépendance comme le tau de Kendall et le rho de Spearman. Sauf indications contraires, les résultats présentés dans ce chapitre ont été obtenus par [19].

2.2 Famille des copules Khi-deux bivariées

2.2.1 Construction de la copule Khi-deux

Soit (Z_1, Z_2), une paire de variables aléatoires de loi Normale de moyennes nulles, de variances unitaires, et de corrélation $\rho \in (-1, 1)$. Pour un vecteur de paramètres de décentralisation $(a_1, a_2) \in \mathbb{R}^2$, la copule Khi-deux bivariée est simplement définie comme la structure de dépendance, c'est-à-dire la copule, de

$$(X_1, X_2) = \left((Z_1 + a_1)^2, (Z_2 + a_2)^2\right).$$

Afin d'obtenir une forme explicite pour cette copule, on note d'abord que les fonctions de répartition marginales de X_1 et X_2 sont respectivement

$$\begin{aligned} G_{a_1}(x) &= \mathbb{P}(X_1 \leq x) = \Phi\left(\sqrt{x} + a_1\right) + \Phi\left(\sqrt{x} - a_1\right) - 1, \\ G_{a_2}(x) &= \mathbb{P}(X_2 \leq x) = \Phi\left(\sqrt{x} + a_2\right) + \Phi\left(\sqrt{x} - a_2\right) - 1, \end{aligned}$$

où $x \geq 0$ et Φ est la fonction de répartition de la loi Normale standard univariée. La copule Khi-deux bivariée, notée C^{χ}_{ρ,a_1,a_2}, correspond donc à la fonction de répartition conjointe de $G_{a_1}(X_1)$ et $G_{a_2}(X_2)$, c'est-à-dire que

$$C^{\chi}_{\rho,a_1,a_2}(u_1, u_2) = \mathbb{P}\left\{G_{a_1}(X_1) \leq u_1, G_{a_2}(X_2) \leq u_2\right\}. \tag{2.1}$$

Soit maintenant la fonction $h_a(u) = \text{sign}(u)\sqrt{G_a^{-1}(|u|)} - a$ définie pour $u \in [-1, 1]$. En partant de l'Équation (2.1) et en se rappelant que $X_1 = (Z_1 + a_1)^2$ et $X_2 = (Z_2 + a_2)^2$,

$$\begin{aligned} C^{\chi}_{\rho,a_1,a_2}(u_1, u_2) &= \mathbb{P}\left\{X_1 \leq G_{a_1}^{-1}(u_1), X_2 \leq G_{a_2}^{-1}(u_2)\right\} \\ &= \mathbb{P}\left\{(Z_1 + a_1)^2 \leq G_{a_1}^{-1}(u_1), (Z_2 + a_2)^2 \leq G_{a_2}^{-1}(u_2)\right\} \\ &= \mathbb{P}\left\{|Z_1 + a_1| \leq \sqrt{G_{a_1}(u_1)}, |Z_2 + a_2| \leq \sqrt{G_{a_2}(u_2)}\right\} \\ &= \mathbb{P}\left\{-\sqrt{G_{a_1}(u_1)} - a_1 \leq Z_1 \leq \sqrt{G_{a_1}(u_1)} - a_1, \right. \\ &\qquad \left. -\sqrt{G_{a_2}(u_2)} - a_2 \leq Z_2 \leq \sqrt{G_{a_2}(u_2)} - a_2\right\} \\ &= \mathbb{P}\left\{h_{a_1}(-u_1) \leq Z_1 \leq h_{a_1}(u_1), h_{a_2}(-u_2) \leq Z_2 \leq h_{a_2}(u_2)\right\} \\ &= \mathbb{P}\left\{Z_1 \leq h_{a_1}(u_1), Z_2 \leq h_{a_2}(u_2)\right\} \\ &\quad - \mathbb{P}\left\{Z_1 \leq h_{a_1}(u_1), Z_2 \leq h_{a_2}(-u_2)\right\} \\ &\quad - \mathbb{P}\left\{Z_1 \leq h_{a_1}(-u_1), Z_2 \leq h_{a_2}(u_2)\right\} \\ &\quad + \mathbb{P}\left\{Z_1 \leq h_{a_1}(-u_1), Z_2 \leq h_{a_2}(-u_2)\right\}. \end{aligned}$$

Enfin, en définissant Φ_ρ comme étant la fonction de répartition de la distribution Normale standard bivariée de corrélation ρ, on obtient l'expression compacte

$$C^{\chi}_{\rho,a_1,a_2}(u_1,u_2) = \sum_{(\epsilon_1,\epsilon_2)\in\{-1,1\}^2} \epsilon_1\epsilon_2\Phi_\rho\left\{h_{a_1}(\epsilon_1 u_1), h_{a_2}(\epsilon_2 u_2)\right\}. \tag{2.2}$$

À partir de l'expression de l'Équation (2.2), on déduit que la densité de la copule Khi-deux peut s'écrire simplement par

$$c^{\chi}_{\rho,a_1,a_2}(u_1,u_2) = h'_{a_1}(u_1)\,h'_{a_2}(u_2) \sum_{(\epsilon_1,\epsilon_2)\in\{-1,1\}^2} \phi_\rho\left\{h_{a_1}(\epsilon_1 u_1), h_{a_2}(\epsilon_2 u_2)\right\},$$

où ϕ_ρ est la densité de la loi Normale standard bivariée. On peut également obtenir une expression alternative pour représenter la copule Khi-deux en fonction de la copule Normale. En effet, en rappelant que la copule Normale bivariée peut s'écrire sous la forme $C^{\mathrm{N}}_\rho(u_1,u_2) = \Phi_\rho\{\Phi^{-1}(u_1),\Phi^{-1}(u_2)\}$, on a

$$\Phi_\rho\left\{h_{a_1}(\epsilon_1 u_1), h_{a_2}(\epsilon_2 u_2)\right\} = C^{\mathrm{N}}_\rho\left\{\Phi\circ h_{a_1}(\epsilon_1 u_1), \Phi\circ h_{a_2}(\epsilon_2 u_2)\right\}.$$

De l'Équation (2.2), on a pour $\widetilde{h}_a(u) = \Phi\circ h_a(u)$ que

$$C^{\chi}_{\rho,a_1,a_2}(u_1,u_2) = \sum_{(\epsilon_1,\epsilon_2)\in\{-1,1\}^2} \epsilon_1\epsilon_2 C^{\mathrm{N}}_\rho\left\{\widetilde{h}_{a_1}(\epsilon_1 u_1), \widetilde{h}_{a_2}(\epsilon_2 u_2)\right\}. \tag{2.3}$$

En dérivant cette dernière expression par rapport à (u_1,u_2), on trouve aussi une représentation alternative de la densité de la copule Khi-deux, à savoir

$$c^{\chi}_{\rho,a_1,a_2}(u_1,u_2) = \sum_{(\epsilon_1,\epsilon_2)\in\{-1,1\}^2} \widetilde{h}'_{a_1}(\epsilon_1 u_1)\widetilde{h}'_{a_2}(\epsilon_2 u_2) c^{\mathrm{N}}_\rho\left\{\widetilde{h}_{a_1}(\epsilon_1 u_1), \widetilde{h}_{a_2}(\epsilon_2 u_2)\right\},$$

où c^{N}_ρ est la densité de la copule Normale bivariée.

2.2.2 Cas particulier où $a_1 \to \infty$ et $a_2 \to \infty$

Tel que mentionné dans le préambule de ce chapitre, la copule Khi-deux est une généralisation de la copule Normale. En effet, lorsque $a_1 \to \infty$ et $a_2 \to \infty$, on retrouve la copule Normale, ce qui veut dire que cette copule est simplement un cas particulier de la famille Khi-deux. Pour le démontrer, il faut d'abord remarquer que

$$\lim_{a\to\infty} \widetilde{h}_a(u) = u\,\mathbb{I}(u \geq 0).$$

Ceci provient du fait que

$$\widetilde{h}_a(u) = \Phi \circ h_a(u) = \mathbb{P}\left\{Z + a \leq \text{sign}(u)\sqrt{G_a^{-1}(|u|)}\right\}.$$

Ainsi, lorsque $a_1 \to \infty$ et $a_2 \to \infty$, tous les termes de l'Équation (2.3) disparaissent, sauf lorsque $\epsilon_1 = \epsilon_2 = 1$. Ainsi, tel qu'anticipé,

$$\lim_{a_1,a_2\to\infty} C^{\chi}_{\rho,a_1,a_2}(u_1,u_2) = C^{\text{N}}_{\rho}(u_1,u_2).$$

2.2.3 Cas particulier où $a_1 = a_2 = 0$

Quand $a_1 = a_2 = 0$, la copule $C^{\chi}_{\rho,0,0}$, simplement notée C^{χ}_{ρ}, s'appelle la copule Khi-deux centrée. Pour obtenir sa forme explicite, on fera appel au lemme suivant.

Lemme 2.1. *On a* $\widetilde{h}_0(u) = (1+u)/2$.

En utilisant la conclusion du Lemme 2.1 dans l'Équation (2.3), on trouve

$$\begin{aligned} C^{\chi}_{\rho}(u_1,u_2) \;=\; & C^{\text{N}}_{\rho}\left(\frac{1+u_1}{2},\frac{1+u_2}{2}\right) - C^{\text{N}}_{\rho}\left(\frac{1+u_1}{2},\frac{1-u_2}{2}\right) \\ & - C^{\text{N}}_{\rho}\left(\frac{1-u_1}{2},\frac{1+u_2}{2}\right) + C^{\text{N}}_{\rho}\left(\frac{1-u_1}{2},\frac{1-u_2}{2}\right). \end{aligned}$$

De plus, comme la copule Normale possède la propriété de symétrie radiale, on a $C_\rho^{\mathrm{N}}(1-u_1, 1-u_2) = u_1 + u_2 - 1 + C_\rho^{\mathrm{N}}(u_1, u_2)$. Quelques calculs directs amènent

$$C_\rho^{\chi}(u_1, u_2) = 2\left\{C_\rho^{\mathrm{N}}\left(\frac{1+u_1}{2}, \frac{1+u_2}{2}\right) - C_\rho^{\mathrm{N}}\left(\frac{1+u_1}{2}, \frac{1-u_2}{2}\right)\right\} - u_2. \tag{2.4}$$

En dérivant cette dernière expression selon u_1 et u_2, on obtient que la densité associée est

$$c_\rho^{\chi}(u_1, u_2) = \frac{1}{2}\left\{c_\rho^{\mathrm{N}}\left(\frac{1+u_1}{2}, \frac{1+u_2}{2}\right) + c_\rho^{\mathrm{N}}\left(\frac{1+u_1}{2}, \frac{1-u_2}{2}\right)\right\}.$$

2.3 Propriétés de la copule Khi-deux bivariée

Après avoir construit la copule Khi-deux pour le cas bivarié et s'être intéressé aux formes sous lesquelles elle peut se présenter, il est primordial de s'attarder sur certaines propriétés de cette famille de copules qui sont très utiles en pratique et qui permettent de simplifier son utilisation. De plus, cela amènera une connaissance et une familiarisation accrues de ces mêmes copules.

En premier lieu, on se rappelle que la copule Khi-deux bivariée possède trois paramètres, à savoir ρ, a_1 et a_2. On peut s'intéresser au comportement de la copule lorsque ces paramètres changent de signe. On trouve alors un cas particulier pertinent qui mérite notre attention dans le lemme qui suit.

Lemme 2.2. *Pour tout $\rho \in (-1, 1), (a_1, a_2) \in \mathbb{R}^2$ et $(u_1, u_2) \in [0, 1]^2$, nous avons*

$$C_{\rho,a_1,a_2}^{\chi}(u_1, u_2) = C_{-\rho,-a_1,a_2}^{\chi}(u_1, u_2) = C_{\rho,-a_1,-a_2}^{\chi}(u_1, u_2).$$

En d'autres mots, la copule Khi-deux est invariante sous le changement de signe d'exactement deux de ses paramètres.

Une conséquence immédiate de ce lemme est le fait que pour la copule Khi-deux bivariée centrée, on a $C^{\chi}_{\rho} = C^{\chi}_{-\rho}$. Alors, dans ce cas précis, nous pouvons restreindre ρ à se situer entre 0 et 1 sans perte de généralité. De façon générale, on dit que la copule C est dominée par la copule D si $C(u_1, u_2) \leq D(u_1, u_2), \forall (u_1, u_2)$. À la lumière de cette dernière remarque, intéressons-nous à la famille des copules Khi-deux centrées. Une propriété intéressante spécifique à cette famille réside dans le fait qu'elle soit ordonnée stochastiquement selon $|\rho| \in [0, 1)$.

Proposition 2.1. *Pour tout* $|\rho| \leq |\rho'| \in [0, 1)$, *nous avons l'inégalité suivante :*

$$C^{\chi}_{\rho}(u_1, u_2) \leq C^{\chi}_{\rho'}(u_1, u_2), \quad \forall (u_1, u_2) \in [0, 1]^2.$$

On sait que toute copule C est bornée par les bornes de Fréchet-Hoeffding W et M correspondant à la dépendance négative et positive parfaite, respectivement. Ces deux copules font partie de la famille des copules Normales car la copule Normale tend vers W lorsque $\rho \to -1$ et vers M lorsque $\rho \to 1$. On dit alors que la famille des copules Normales est complète. Afin d'étudier les copules Khi-deux sous cet angle, on considérera d'abord la copule Khi-deux centrée. On note d'abord que si $\rho = 0$, alors de l'Équation (2.4), on déduit $C^{\chi}_{0} = \Pi(u_1, u_2)$. Similairement, on montre que $\lim_{\rho \to 1} C^{\chi}_{\rho}(u_1, u_2) = M(u_1, u_2)$. À partir de la Proposition 2.1, on peut établir que pour tout $(u_1, u_2) \in [0, 1]^2$, $\Pi(u_1, u_2) \leq C^{\chi}_{\rho}(u_1, u_2) \leq M(u_1, u_2)$. Autrement dit, la famille des copules Khi-deux centrées ne permet que de la dépendance positive.

Dans le cas général d'un paramètre de décentralisation (a_1, a_2) non-nul, les choses se compliquent quelque peu. Lorsque $\rho = 0$, on obtient toujours que la copule Khi-deux $C^{\chi}_{\rho, a_1, a_2}$ correspond à la copule d'indépendance peu importe les valeurs de a_1 et a_2. En revanche, lorsque $\rho \to 1$, on a $Z_1 = Z_2$ presque sûrement dans la représentation de la copule Khi-deux. En posant $x \vee y = \max(x, y)$ et $x \wedge y = \min(x, y)$ et en se

remémorant le fait que $\widetilde{h}_a(u) = \Phi \circ h_a(u)$, on montre alors que

$$\begin{aligned}\lim_{\rho\to 1} C^{\chi}_{\rho,a_1,a_2}(u_1,u_2) &= \mathbb{P}\left\{h_{a_1}(-u_1) \vee h_{a_2}(-u_2) \leq Z_1 \leq h_{a_1}(u_1) \wedge h_{a_2}(u_2)\right\} \\ &= \max\left\{0, \left(\widetilde{h}_{a_1}(u_1) \wedge \widetilde{h}_{a_2}(u_2)\right) - \left(\widetilde{h}_{a_1}(-u_1) \vee \widetilde{h}_{a_2}(-u_2)\right)\right\}.\end{aligned}$$

Cette dernière expression n'est pas, en général, la borne supérieure de Fréchet-Hoeffding.

Proposition 2.2. *Pour tout $u \in [0,1]$, on a $\widetilde{h}_a(u) - \widetilde{h}_a(-u) = u$.*

En utilisant ce dernier résultat, on peut déduire que

$$\begin{aligned}\lim_{\rho\to 1} C^{\chi}_{\rho,a,a}(u_1,u_2) &= \max\left\{0, \left(\widetilde{h}_a(u_1) \wedge \widetilde{h}_a(u_2)\right) - \left(\widetilde{h}_a(-u_1) \vee \widetilde{h}_a(-u_2)\right)\right\} \\ &= \max\left\{0, \widetilde{h}_a(u_1 \wedge u_2) - \widetilde{h}_a(-(u_1 \wedge u_2))\right\} \\ &= \min(u_1,u_2).\end{aligned}$$

Enfin, pour traiter du cas $\rho \to -1$, on peut invoquer le Lemme 2.2 et utiliser le fait que la copule Khi-deux bivariée est invariante sous le changement de signe de deux de ses paramètres pour écrire

$$\lim_{\rho\to -1} C^{\chi}_{\rho,a_1,a_2} = \lim_{\rho\to 1} C^{\chi}_{\rho,a_1,-a_2}.$$

Une propriété importante des copules en général réside dans la symétrie (ou l'asymétrie) de celles-ci. Certaines copules présentent ce qu'on appelle de la symétrie diagonale, c'est-à-dire que $C(u_1,u_2) = C(u_2,u_1)$ poru tout $(u_1,u_2) \in [0,1]^2$. C'est le cas notamment de la copule Normale. C'est aussi le cas de la copule Khi-deux bivariée lorsque le paramètre de décentralisation est tel que $a_1 = a_2 = a \in \mathbb{R}$. On obtient alors que $C^{\chi}_{\rho,a,a}(u_1,u_2) = C^{\chi}_{\rho,a,a}(u_2,u_1)$ et ce, pour n'importe quel $u_1,u_2 \in [0,1]$. Ceci peut être déduit directement de l'Équation (2.3) en utilisant le fait que la copule Normale soit symétrique diagonalement. La même situation se produit dans le cas où nous avons $a_1 = -a_2$, conséquence du Lemme 2.2. Il est toutefois beaucoup plus intéressant de s'attarder à l'asymétrie des copules Khi-deux. On se rappelle que cette famille de

copules permet de modéliser plusieurs structures de dépendance. Cette richesse réside dans la génération d'un très grand nombre de modèles asymétriques, avantage considérable par rapport à la famille des copules Normales. Donc, le cas général $|a_1| \neq |a_2|$ renferme plusieurs situations pertinentes et pratiques.

À l'aide du fait que $\widetilde{h}_a(u) \to u\mathbb{I}(u \geq 0)$ lorsque $a \to \infty$ et du Lemme 2.1, il nous est possible de trouver l'expression de copules Khi-deux pour des valeurs spécifiques des deux composantes du paramètre de décentralisation. Par exemple, dans le cas où $(a_1, a_2) = (a, \infty)$, on obtient la copule

$$C^{\chi}_{\rho,a,\infty}(u_1, u_2) = C^{\mathrm{N}}_{\rho}\left\{u_1, \widetilde{h}_a(u_2)\right\} - C^{\mathrm{N}}_{\rho}\left\{u_1, \widetilde{h}_a(-u_2)\right\}. \tag{2.5}$$

Une mesure d'asymétrie a été introduite par [17]. Pour une copule Khi-deux C^{χ}_{ρ,a_1,a_2}, cette mesure s'écrit

$$\delta\left(C^{\chi}_{\rho,a_1,a_2}\right) = 3 \sup_{(u_1,u_2)\in[0,1]^2} \left|C^{\chi}_{\rho,a_1,a_2}(u_1, u_2) - C^{\chi}_{\rho,a_1,a_2}(u_2, u_1)\right|.$$

En faisant quelques observations, il est facile de voir que le niveau d'asymétrie augmente au fur et à mesure que l'écart $|a_1| - |a_2|$ se creuse en valeur absolue. Donc, en ce qui a trait au paramètre de décentralisation, la copule Khi-deux la plus asymétrique est celle présentée dans le cas où $a_1 \to \infty$ et $a_2 = 0$. En remplaçant a par 0 dans l'Équation (2.5), on obtient la copule via

$$C^{\chi}_{\rho,0,\infty}(u_1, u_2) = C^{\mathrm{N}}_{\rho}\left(u_1, \frac{1+u_2}{2}\right) - C^{\mathrm{N}}_{\rho}\left(u_1, \frac{1-u_2}{2}\right).$$

On peut aussi noter que la mesure d'asymétrie δ est une fonction croissante par rapport à la valeur absolue de la corrélation ρ et ce, peu importe la valeur du paramètre de décentralisation. Alors, pour ce paramètre, la copule Khi-deux la plus asymétrique est obtenue lorsque $|\rho| \to 1$. En combinant les observations faites sur le sujet, il est possible de trouver la copule Khi-deux présentant réellement le plus grand niveau

d'asymétrie. On obtient la forme qui suit pour la copule sous ces conditions

$$\begin{aligned}\lim_{|\rho|\to 1} C^{\chi}_{\rho,0,\infty}(u_1,u_2) &= \max\left\{0, \left(\widetilde{h}_0(u_1)\wedge\widetilde{h}_\infty(u_2)\right) - \left(\widetilde{h}_0(-u_1)\vee\widetilde{h}_\infty(-u_2)\right)\right\}\\ &= \max\left\{0, \left(\frac{1+u_1}{2}\wedge u_2\right) - \left(\frac{1+u_1}{2}\vee -u_2\right)\right\}\\ &= \max\left\{0, \left(\frac{1+u_1}{2}\wedge u_2\right) - \frac{1+u_1}{2}\right\}.\end{aligned}$$

2.4 Mesures de dépendance

2.4.1 Opérateur de concordance

Le Chapitre 1 a introduit des mesures de dépendance et montré leurs expressions dans le cas de la copule Normale. L'idée ici est d'obtenir de telles expressions pour la famille des copules Khi-deux. À cette fin, on établit d'abord un résultat concernant l'opérateur de concordance.

Proposition 2.3. *Pour $\rho, \widetilde{\rho} \in (-1,1)$, l'opérateur de concordance de la copule Khi-deux est donné pour $b_1 = -\sqrt{2}a_1$ et $b_2 = -\sqrt{2}a_2$ par*

$$\begin{aligned}Q\left(C^{\chi}_{\rho,a_1,a_2}, C^{\chi}_{\widetilde{\rho},a_1,a_2}\right) &= 16\,\Phi^4_\Sigma(0,0,b_1,b_2)\\ &\quad - 8\left\{\Phi^3_{\Sigma'}(0,0,b_1) + \Phi^3_{\Sigma'}(0,0,b_2) + \Phi^3_{\Sigma'}(b_1,b_2,0) + \Phi^3_{\Sigma'}(b_2,b_1,0)\right\}\\ &\quad + 4\left\{\Phi_{\rho^+}(0,0) + \Phi_{\rho^+}(b_1,b_2) + \Phi_{\rho^-}(0,b_1) + \Phi_{\rho^-}(0,b_2)\right\} - 1,\end{aligned}$$

où $\rho^+ = (\rho+\widetilde{\rho})/2$, $\rho^- = (\rho-\widetilde{\rho})/2$ et Φ^4_Σ est la fonction de répartition de la distribution Normale standard en quatre dimensions avec pour matrice de corrélation $(\Sigma_{ij})_{i,j=1}^4$ telle que $\Sigma_{21} = \Sigma_{43} = \rho^+, \Sigma_{32} = \Sigma_{41} = \rho^-$ et $\Sigma_{31} = \Sigma_{42} = 0$. De plus, $\Phi^3_{\Sigma'}$ est la fonction de répartition de la distribution Normale standard en trois dimensions ayant pour matrice de corrélation Σ' telle que $\Sigma'_{ij} = \Sigma_{ij}, i,j \in \{1,2,3\}$.

Lorsque $\rho = \widetilde{\rho} = 0$, l'opérateur de concordance est nul car $C^{\chi}_{\rho,a_1,a_2} = C^{\chi}_{\widetilde{\rho},a_1,a_2} = \Pi$. À partir de la Proposition 2.3, on voit bien que dans le cas spécifique où $a_1 = a_2 \to \infty$, on a $Q(C^{\chi}_{\rho,a_1,a_2}, C^{\chi}_{\widetilde{\rho},a_1,a_2}) \to 4\Phi_{\rho^+}(0,0) - 1$; ceci concorde avec l'expression de l'opérateur de concordance pour la copule Normale bivariée. L'expresssion de l'opérateur de concordance dans le cas des copules Khi-deux centrées est donnée dans ce qui suit.

Corollaire 2.1. *Pour Σ, ρ^+ et ρ^- définis à la Proposition 2.3, l'opérateur de concordance de la copule Khi-deux bivariée centrée est donné par* $Q(C^{\chi}_{\rho}, C^{\chi}_{\widetilde{\rho}}) = 16\,\Phi^4_{\Sigma}(0,0,0,0) - 8\Phi_{\rho^+}(0,0) - 8\Phi_{\rho^-}(0,0) + 3$.

2.4.2 Tau de Kendall de la copule Khi-deux

On commence par énoncer un résultat général concernant la copule Khi-deux bivariée.

Proposition 2.4. *Soit $\tau(C^{\mathrm{N}}_{\rho}) = (2/\pi)\sin^{-1}\rho$, le tau de Kendall de la copule Normale. On définit le tau de Kendall de la copule Khi-deux par :*

$$\tau(C^{\chi}_{\rho,a_1,a_2}) = \tau(C^{\mathrm{N}}_{\rho})\left\{4\Phi_{\rho}(\sqrt{2}a_1, \sqrt{2}a_2) - 2\Phi(\sqrt{2}a_1) - 2\Phi(\sqrt{2}a_2) + 1\right\}.$$

En particulier, le tau de Kendall de la copule Khi-deux symétrique $C^{\chi}_{\rho,a,a}$ est

$$\tau(C^{\chi}_{\rho,a,a}) = \tau(C^{\mathrm{N}}_{\rho})\left\{4\Phi_{\rho}(\sqrt{2}a, \sqrt{2}a) - 4\Phi(\sqrt{2}a) + 1\right\}. \tag{2.6}$$

À partir de la Proposition 2.4, on obtient facilement que $\lim_{a\to\infty}\tau(C^{\chi}_{\rho,a,a}) = \tau(C^{\mathrm{N}}_{\rho})$, en cohérence avec le fait que la copule Normale apparaît comme cas particulier de la copule Khi-deux générale lorsque $a_1 \to \infty$ et $a_2 \to \infty$. Le résultat suivant donne une formule pour le tau de Kendall de la copule Khi-deux centrée.

Corollaire 2.2. *On a $\tau(C^{\chi}_{\rho}) = \{\tau(C^{\mathrm{N}}_{\rho})\}^2 = \{(2/\pi)\sin^{-1}\rho\}^2$.*

Lorsque $\rho \to 1$, la Proposition 2.4 permet d'obtenir

$$\lim_{\rho \to 1} \tau(C^{\chi}_{\rho,a_1,a_2}) = 1 - 2\left\{\Phi\left(\sqrt{2}\max(a_1,a_2)\right) - \Phi\left(\sqrt{2}\min(a_1,a_2)\right)\right\}.$$

En effectuant quelques petits raisonnements, on remarque que la précédente limite est égale à 1 si et seulement si $a_1 = a_2$. Dans le cas où $a_1 = -a_2 = a > 0$, cette limite prend alors la valeur $3 - 4\Phi(\sqrt{2}a)$, ce qui implique que le tau de Kendall est négatif lorsque $a > \Phi^{-1}(3/4)/\sqrt{2} \approx 0,477$. En particulier, $\lim_{\rho \to 1} \tau(C^{\chi}_{\rho,\infty,-\infty}) = -1$. On trouve aussi un tau de Kendall négatif lorsque $a_1 < 0$ et $a_2 \to \infty$, car alors $\tau(C^{\chi}_{\rho,a_1,\infty}) = \tau(C^{\mathrm{N}}_{\rho})\{2\Phi(\sqrt{2}a_1) - 1\} < 0$.

2.4.3 Rho de Spearman de la copule Khi-deux

Du Chapitre 1, on sait que $\rho_S(C) = 3\,Q(C,\Pi)$, où Π est la copule d'indépendance. Puisque $C^{\chi}_{0,a_1,a_2} = \Pi$, on peut donc invoquer la Proposition 2.3 et déduire que

$$\rho_S(C^{\chi}_{\rho,a_1,a_2}) = 3\,Q\left(C^{\chi}_{\rho,a_1,a_2}, C^{\chi}_{0,a_1,a_2}\right).$$

En général, cette expression n'a pas de forme explicite. On obtient néanmoins une formule compacte dans le cas de la copule Khi-deux centrée, à savoir

$$\rho_S(C^{\chi}_{\rho}) = 48\left\{\Phi^4_{\Sigma}(0,0,0,0) - \Phi_{\rho/2}(0,0)\right\} + 9,$$

où la matrice Σ est telle que $\Sigma_{21} = \Sigma_{43} = \Sigma_{32} = \Sigma_{41} = \rho/2$ et $\Sigma_{31} = \Sigma_{42} = 0$.

2.5 Copules Khi-deux multidimensionnelles

2.5.1 Construction du modèle

Les modèles spatio-temporels qui seront développés au chapitre suivant seront basés sur la copule Khi-deux multidimensionnelle ; celle-ci est, d'une certaine manière, une extension du modèle développé dans le cas bidimensionnel. Pour la construire, soit une matrice de corrélation Σ dont toutes les entrées sont non-négatives, ainsi qu'un paramètre de décentralisation $a \in \mathbb{R}$. Ensuite, à partir d'un vecteur aléatoire $(Z_1, \ldots, Z_d) \sim \Phi_\Sigma$, on définit la copule Khi-deux d-dimensionnelle $C^\chi_{\Sigma,a}$ comme étant la structure de dépendance de $(X_1, \ldots, X_d)$ où $X_j = (Z_j + a)^2$ pour chaque $j \in \{1, \ldots, d\}$.

Dans leur article, [19] montrent qu'une expression pour la copule Khi-deux est

$$C^\chi_{\Sigma,a}(u_1, \ldots, u_d) = \sum_{(\epsilon_1,\ldots,\epsilon_d)\in\Xi} \left(\prod_{j=1}^{d} \epsilon_j\right) \Phi_\Sigma \left\{h_a(\epsilon_1 u_1), \ldots, h_a(\epsilon_d u_d)\right\}, \tag{2.7}$$

où $\Xi = \{-1, +1\} \times \cdots \times \{-1, +1\}$. En dérivant cette expression, on déduit qu'une expression pour la densité de la copule Khi-deux est

$$c^\chi_{\Sigma,a}(u_1, \ldots, u_d) = \left(\prod_{j=1}^{d} h'_a(u_j)\right) \sum_{(\epsilon_1,\ldots,\epsilon_d)\in\Xi} \phi_\Sigma \left\{h_a(\epsilon_1 u_1), \ldots, h_a(\epsilon_d u_d)\right\}. \tag{2.8}$$

On peut déduire une forme pour la copule Khi-deux multidimensionnelle en fonction de la copule Normale en utilisant la relation connue entre cette copule et la fonction de répartition de la loi Normale d-dimensionnelle Φ_Σ. En effet, [19] établissent que

$$C^\chi_{\Sigma,a}(u_1, \ldots, u_d) = \sum_{(\epsilon_1,\ldots,\epsilon_d)\in\Xi} \left(\prod_{j=1}^{d} \epsilon_j\right) C^{\mathrm{N}}_\Sigma \left\{\widetilde{h}_a(\epsilon_1 u_1), \ldots, \widetilde{h}_a(\epsilon_d u_d)\right\}. \tag{2.9}$$

La copule Normale est un cas particulier de la famille Khi-deux quand $a \to \infty$. En

effet, puisque $\widetilde{h}_a(u) \to u\,\mathrm{I}(u \geq 0)$ lorsque $a \to \infty$, tous les termes de la sommation dans l'Équation (2.9) sont nuls, sauf le cas où $\epsilon_1 = \cdots = \epsilon_d = 1$. On a donc

$$\lim_{a\to\infty} C^{\chi}_{\Sigma,a}(u_1, \ldots, u_d) = C^{\mathrm{N}}_{\Sigma}(u_1, \ldots, u_d).$$

Lorsque Σ est la matrice identité I_d, $C^{\chi}_{I_d,a}(u_1, \ldots, u_d) = u_1 \times \cdots \times u_d$, c'est-à-dire la copule de l'indépendance d-dimensionnelle. Ce résultat découle directement du l'Équation (2.9). Ainsi, en utilisant le fait que $C^{\mathrm{N}}_{I_d}$ est la copule d'indépendance, on a

$$C^{\chi}_{I_d,a}(u_1, \ldots, u_d) = \prod_{j=1}^{d} \left\{ \widetilde{h}_a(u_j) - \widetilde{h}_a(-u_j) \right\} = u_1 \times \cdots \times u_d,$$

où on a utilisé le fait que $\widetilde{h}_a(u) - \widetilde{h}_a(-u) = u$, tel qu'établi à la Proposition 2.2.

2.5.2 Cas particuliers

Dans un procédé analogue au cas bivarié, il est pertinent de s'attarder à certaines conditions spécifiques sous lesquelles la copule Khi-deux multivariée se transforme en copules qui nous sont familiaires. Un premier exemple d'une telle situation survient lorsque nous supposons que la matrice de corrélation est telle que $\Sigma_{jj'} \to 1$ pour tout $j, j' \in \{1, \ldots, d\}$. Après quelques petites observations, on remarque que dans cette éventualité, la copule Normale multivariée correspond à la borne supérieure de Fréchet-Hoeffding multivariée qui s'exprime sous la forme $M(\mathbf{u}) = \min_{j\in\{1,\ldots,d\}} u_j$. De plus, on sait que dans la représentation $((Z_1 + a)^2, \ldots, (Z_d + a)^2)$, nous avons $Z_1 = \cdots = Z_d$ presque sûrement pour la condition sur la matrice de corrélation mentionnée. Nous sommes alors en mesure d'exprimer la copule $M(\mathbf{u})$ de la façon

suivante toujours en faisant usage de la propriété de la Proposition 2.2 :

$$\begin{aligned} C^{\chi}_{\Sigma,a}(\mathbf{u}) &= \max\left\{0, \min_{j\in\{1,\ldots,d\}} \widetilde{h}_a(u_j) - \max_{j\in\{1,\ldots,d\}} \widetilde{h}_a(-u_j)\right\} \\ &= \widetilde{h}_a\left\{M(\mathbf{u})\right\} - \widetilde{h}_a\left\{-M(\mathbf{u})\right\} \\ &= M(\mathbf{u}). \end{aligned}$$

Nous avons aussi vu que dans la cas bivarié, la copule Khi-deux centrée présentait certaines particularités étonnantes. Il est envisageable de pouvoir retrouver cette copule dans une situation multivariée. Par un raisonnement relativement intuitif, il est évident que cette copule se présente lorsque le paramètre de décentralisation a est nul. Reprenant la forme de la copule Khi-deux multivariée explicitée à l'Équation (2.9) ainsi que le résultat du Lemme 2.1, on peut déterminer que la forme de la copule Khi-deux centrée multivariée est donnée via l'expression

$$C^{\chi}_{\Sigma,0}(\mathbf{u}) = \sum_{\epsilon\in\Xi}\left(\prod_{j=1}^{d}\epsilon_j\right) C^{\mathrm{N}}_{\Sigma}\left(\frac{1+\epsilon_1 u_1}{2}, \ldots, \frac{1+\epsilon_d u_d}{2}\right).$$

Une simple dérivation par rapport à chacune des variables de cette formule nous permet de trouver la densité de cette copule. On obtient

$$c^{\chi}_{\Sigma,0}(\mathbf{u}) = \frac{1}{2^d}\sum_{\epsilon\in\Xi} c^{\mathrm{N}}_{\Sigma}\left(\frac{1+\epsilon_1 u_1}{2}, \ldots, \frac{1+\epsilon_d u_d}{2}\right).$$

2.5.3 Restrictions sur la matrice de Kendall

Un aspect important dans l'étude des copules est la détermination des mesures de dépendance. Nous avons obtenu des résultats intéressants sur ce sujet notamment pour l'opérateur de concordance et le tau de Kendall. Nous verrons désormais ce qu'il en est pour la copule Khi-deux multivariée. Dans ce cas précis, le tau de Kendall est remplacé par une matrice de Kendall qui est composée des tau de Kendall bivariés associés à chacune des paires de variables aléatoires différentes composant la copule.

Certaines contraintes doivent être imposées sur cette matrice de Kendall pour que la structure de dépendance à l'étude puisse être modélisée par une famille de copules donnée. Par exemple, prenons une copule arbitraire C tridimensionnelle avec les tau de Kendall bivariés associés τ_{12}, τ_{13} et τ_{23}. Alors [8] a démontré les contraintes générales $\tau_{12}, \tau_{13} \in [-1, 1]$ et $-1 + |\tau_{12} + \tau_{13}| \leq \tau_{23} \leq 1 - |\tau_{12} - \tau_{13}|$. À noter que les bornes inférieures et supérieures sont atteintes lorsque la copule C est la copule Normale. Donc, les éléments de la matrice de Kendall de la copule C tridimensionnelle doivent respecter ces conditions pour qu'elle puisse être utiliser dans la modélisation d'une structure de dépendance quelconque. Le prochain résultat nous donne les restrictions sur les tau de Kendall bivariés pour la copule Khi-deux tridimensionnelle.

Proposition 2.5. *Soient τ_{12}, τ_{13} et τ_{23}, les tau de Kendall par paires de la copule Khi-deux en trois dimensions avec paramètre de décentralisation $a \geq 0$. Alors, $\tau_{12} \in [-1, 1]$, $\tau_{13} \in [-1, 1]$ et $\tau_{23} \in [g_a(\gamma_1 - \sqrt{\gamma_2}), g_a(\gamma_1 + \sqrt{\gamma_2})]$, où*

$$\begin{aligned} g_a(x) &= \tau(C_x^{\mathrm{N}}) \left\{ 4\,\Phi_x\left(\sqrt{2}a, \sqrt{2}a\right) - 4\Phi\left(\sqrt{2}a\right) + 1 \right\}, \\ \gamma_1 &= g_a^{-1}(\tau_{12})\, g_a^{-1}(\tau_{13}), \\ \gamma_2 &= \left\{1 - \left(g_a^{-1}(\tau_{12})\right)^2\right\} \left\{1 - \left(g_a^{-1}(\tau_{13})\right)^2\right\}. \end{aligned}$$

Il est important de noter que pour $a \geq 0$ quelconque et des tau de Kendall τ_{12}, τ_{13} et τ_{23} arbitraires mais respectant toujours les conditions de la Proposition 2.5, on peut trouver une matrice de corrélation Σ_a de dimension 3 qui est telle que les tau de Kendall de la matrice Khi-deux $C_{\Sigma_a, a}$ correspondent à τ_{12}, τ_{13} et τ_{23}. Les éléments hors de la diagonale de cette matrice Σ_a sont donnés par la formule $g_a^{-1}(\tau_{jj'})$ pour $j \neq j' \in \{1, 2, 3\}$. Le corollaire suivant s'intéresse maintenant à ce qu'il advient lorsque la copule Khi-deux multivariée est centrée.

Corollaire 2.3. *Soient τ_{12}, τ_{13} et τ_{23}, les tau de Kendall bivariés de la copule Khi-deux centrée de dimension 3. Alors, $\tau_{12}, \tau_{13} \in [0, 1]$ et*

$$\left\{\max\left(0, \sqrt{\tau_{12}} + \sqrt{\tau_{13}} - 1\right)\right\}^2 \leq \tau_{23} \leq \left(1 - |\sqrt{\tau_{12}} - \sqrt{\tau_{13}}|\right)^2.$$

2.6 Estimation des paramètres de la copule Khi-deux

Avant d'utiliser une copule dans une situation concrète, principalement pour la modélisation spatiale, il est important de remarquer que plusieurs paramètres de celle-ci sont inconnus. Il est essentiel d'estimer ces paramètres pour être en mesure d'utiliser cette copule. Nous allons donc présenter deux méthodes d'estimation des paramètres de la copule Khi-deux multivariée, à savoir (i) un maximum de vraisemblance basé sur les rangs et (ii) une version de l'estimateur en (i) basé uniquement sur les paires. Avant de décrire ces méthodes, nous allons présenter un bref rappel sur la méthode générale de l'estimation par maximum de vraisemblance. À noter que l'efficacité et la précision de ces estimateurs ont été étudiées par [19] à l'aide de simulations, et ce sous plusieurs scénarios présentant des caractéristiques distinctes.

2.6.1 Rappel sur la méthode du maximum de vraisemblance

Tout d'abord, nous devons déterminer les paramètres à estimer. Si on se ramène à la construction de cette copule, nous avons introduit la matrice de corrélation Σ et le paramètre de décentralisation a. Les deux premières méthodes pour estimer ces paramètres sont basées sur l'expression de la fonction de vraisemblance qui prend la forme d'une fonction de probabilités conditionnelles décrivant les valeurs d'une loi statistique en fonction des paramètres supposés connus. D'un point de vue mathématique, cette fonction est donnée par

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_\theta(X_i),$$

où f_θ est la densité de probabilités de F_θ, la distribution marginale selon laquelle sont distribuées les variables aléatoires indépendantes $X_1, \ldots, X_n$. Souvent, il est plus

facile de travailler avec la fonction de log-vraisemblance donnée par

$$\mathcal{L}^*(\theta) = \ln \mathcal{L}(\theta) = \sum_{i=1}^{n} \ln f_\theta(X_i).$$

L'estimateur du maximum de vraisemblance est alors la valeur de $\theta \in \Theta$ qui maximise la fonction $\mathcal{L}^*(\theta)$. On a alors

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}^*(\theta).$$

2.6.2 Estimateurs à maximum de vraisemblance

Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$, où $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$, des copies indépendantes du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$ de façon à ce que $\mathbb{P}(\mathbf{X} \leq \mathbf{x}) = C^{\chi}_{\Sigma,a}\{F_1(x_1), \ldots, F_d(x_d)\}$ avec $\mathbf{x} = (x_1, \ldots, x_d)$. Il est important de mentionner que les approches mises en application sont qualifiées de semi-paramétriques car nous supposons que les distributions marginales $F_1, \ldots, F_d$ sont inconnues. Alors, ces distributions seront approximées respectivement par les fonctions empiriques rééchelonnées $F_{n1}, \ldots, F_{nd}$ définies pour chaque $j \in \{1, \ldots, d\}$ comme étant

$$F_{nj}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}\left(X_{ij} \leq x\right).$$

La raison pour laquelle on divise la somme par la quantité $n+1$ plutôt que par n réside dans le fait que quelques petits problèmes numériques peuvent se glisser dans les calculs des estimateurs si tel n'est pas le cas. En utilisant les fonctions empiriques rééchelonnées, on crée une méthode basée sur les rangs des observations. En effet, le calcul de $(n+1)F_{nj}(X_{ij})$ donne le rang de X_{ij} parmi la suite $X_{1j}, \ldots, X_{nj}$. Cette configuration nous permet d'obtenir un premier estimateur pour nos deux paramètres, appelé estimateur de pseudo-vraisemblance (PL), qui s'exprime de la façon suivante,

où $\mathcal{M}$ est l'espace des matrices de corrélation dont les entrées sont non-négatives :

$$\left(\widehat{\Sigma}_n^{\mathrm{PL}}, \widehat{a}_n^{\mathrm{PL}}\right) = \underset{(\Sigma,a)\in\mathcal{M}\times\mathbb{R}^+}{\operatorname{argmax}} \sum_{i=1}^{n} \ln c_{\Sigma,a}^{\chi}\left\{F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id})\right\},$$

où $c_{\Sigma,a}^{\chi}$ est la densité de la copule Khi-deux multivariée. Ce type d'estimateur a été étudié entre autres par [5].

2.6.3 Estimateurs à maximum de vraisemblance par paires

L'estimateur à pseudo-vraisemblance n'est pas adapté à toutes les situations. Comme celui-ci fait appel à la fonction de densité de la copule Khi-deux multivariée et que cette dernière devient de plus en plus complexe avec l'augmentation du nombre de dimensions du problème, les calculs l'impliquant deviennent difficiles et laborieux. Pour remédier à ce problème, la deuxième méthode d'estimation des paramètres nous indique d'utiliser la vraisemblance par paires, qui n'est rien de plus qu'un cas spécial de la vraisemblance composée détaillée dans l'ouvrage [22]. Alors, pour chaque $j < j' \in \{1, \ldots, d\}$, soit $c_{\rho_{jj'},a}^{\chi}$, la densité associée à la copule bivariée formée à partir de la paire aléatoire $(X_j, X_{j'})$. On obtient alors un nouvel estimateur, appelé estimateur de pseudo-vraisemblance par paires (Pair), qui est donné par l'expression

$$\left(\widehat{\Sigma}_n^{\mathrm{Pair}}, \widehat{a}_n^{\mathrm{Pair}}\right) = \underset{(\Sigma,a)\in\mathcal{M}\times\mathbb{R}^+}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ \sum_{j<j'\in\{1,\ldots,d\}} \ln c_{\rho_{jj'},a}^{\chi}\{F_{nj}(X_{ij}), F_{nj'}(X_{ij'})\} \right\}.$$

On rappelle que $\mathcal{M}$ est l'espace des matrices de corrélation dont les entrées sont non-négatives.

Chapitre 3

Nouveaux modèles spatio-temporels basés sur la copule Khi-deux

3.1 Motivation et contexte général

On en arrive maintenant à la contribution principale de ce mémoire, à savoir la construction de modèles statistiques pour l'analyse de données qui possèdent à la fois une composante spatiale et une composante temporelle. Les modèles proposés seront basés sur la famille des copules Khi-deux. Ces structures de dépendance fournissent une alternative intéressante à la copule Normale, notamment parce qu'elles sont en mesure de modéliser en grandes dimensions et que contrairement à la copule Normale, elles permettent l'asymétrie radiale. Également, une copule Khi-deux multidimensionnelle se paramétrise en fonction des paires via une matrice de corrélation, ce qui est une condition essentielle dans un contexte de modélisation spatiale.

Dans la littérature scientifique récente, la modélisation spatiale à l'aide de copules s'est effectuée dans un cadre où les observations à diverses stations étaient, d'une certaine

manière, fixées dans le temps. Les méthodes proposées dans la littérature supposent ainsi qu'une seule réplique par site d'échantillonnage est disponible. Les nouveaux modèles proposés dans ce mémoire ont l'originalité de permettre l'inclusion d'une composante temporelle, c'est-à-dire qu'ils pourront traiter de cas où une succession de données pour chaque site est disponible. Ainsi, les observations à un temps donné pourront dépendre du passé, ce que l'on appelle la dépendance sérielle. Autrement dit, on suppose un champ aléatoire X qui est observé à d sites et ce de façon successive à n intervalles de temps également espacés. On va ainsi construire des modèles où la loi conjointe de ce champ aléatoire, peu importe le moment dans le temps, aura une structure de dépendance caractérisée par une copule Khi-deux d-dimensionnelle telle que définie à l'Équation (2.8). Pour ce faire, quatre situations qui nécessitent un traitement spécifique seront à distinguer, à savoir les cas où les données

(i) sont sériellement indépendantes et les marges sont continues ;

(ii) sont sériellement indépendantes et les marges sont discontinues en zéro ;

(iii) sont sériellement dépendantes et les marges sont continues ;

(iv) sont sériellement dépendantes et les marges sont discontinues en zéro.

Une fois que ces modèles seront construits, des stratégies pour l'estimation des différents paramètres seront proposées en adoptant les lignes de conduite suivantes :

— Parce que typiquement, il y aura un grand nombre d de sites, on estimera toujours, en premier lieu, les paramètres marginaux associés à chacun des sites ;

— Ensuite, conditionnellement à l'estimation des paramètres marginaux effectuée à l'étape précédente, on estimera les paramètres de la dépendance spatiale ;

— Pour l'estimation des paramètres spatiaux, on va privilégier une approche de vraisemblance par paires, car la pleine vraisemblance d-dimensionnelle de la copule Khi-deux possède 2^d termes, ce qui est très lourd à implémenter.

Avant de procéder à la description des nouveaux modèles, la section suivante offre une revue concernant l'usage des copules pour le traitement de données spatiales.

3.2 Utilisation des copules en statistique spatiale

3.2.1 Généralités sur la statistique spatiale

La statistique spatiale permet de décrire et de modéliser des données ayant un référent dans l'espace. Cette branche des sciences statistiques trouve des applications dans divers domaines comme la météorologie, l'hydrologie et la géologie. Ce qui sert généralement de base à la modélisation spatiale est le concept de champ aléatoire défini sur une certaine région $S \subseteq \mathbb{R}^2$. Les informations sur ce champ aléatoire sont fragmentaires au sens où elles sont disponibles seulement à un nombre limité de sites, et non pour toute la région S.

L'utilité principale de la modélisation spatiale est la possibilité de prédire à un lieu qui n'a pas été échantillonné, ce qui s'appelle faire de l'interpolation spatiale. Une méthode d'interpolation spatiale devenue classique a été introduite par [12]. Cet auteur a d'ailleurs donné le nom à cette technique : le krigeage. Plusieurs autres méthodes ont été développées au fil du temps ; pour des détails, voir [20] et [4].

Formellement, soit un champ aléatoire $\{X(\mathbf{s}); \mathbf{s} \in S \subseteq \mathbb{R}^2\}$ pour lequel on dispose de l'information concernant d sites d'observations $\mathbf{s}_1, \ldots, \mathbf{s}_d \in S \subseteq \mathbb{R}^2$. On connaît aussi les distances qui séparent ces sites, stockées dans une matrice symétrique $\Delta \in \mathbb{R}^{d \times d}$ telle que $\Delta_{jj'} = \|\mathbf{s}_j - \mathbf{s}_{j'}\|$, où $\|\cdot\|$ est la norme euclédienne. Le champ aléatoire X évalué aux sites d'échantillonnage est noté $X_1, \ldots, X_d$, où $X_j = X(\mathbf{s}_j)$ pour chaque $j \in \{1, \ldots, d\}$. Pour des fins de modélisation et de prédiction, on s'intéresse à la distribution conjointe du vecteur aléatoire $(X_1, \ldots, X_d)$. On émettra l'hypothèse fondamentale que le champ aléatoire X est isotrope, ce qui signifie que le niveau de dépendance entre deux sites dépend uniquement de la distance qui les sépare.

3.2.2 Fonctions de lien

Sous l'hypothèse d'isotropie, on souhaite moduler le niveau de dépendance entre deux stations en tenant compte de la distance qui les sépare. À cette fin, on utilise une fonction de lien $g : [0, \infty] \to [0, 1]$ qui satisfait les conditions suivantes :

- $(\mathcal{C}_1)$ g est décroissante ;
- $(\mathcal{C}_2)$ $g(0) = 1$;
- $(\mathcal{C}_3)$ $g(x) \to 0$ quand $x \to \infty$.

À partir d'une fonction de lien g donnée, on définit ensuite une matrice de corrélation Γ_θ telle que pour un certain $\theta > 0$, son élément à la position (j, j') est

$$(\Gamma_\theta)_{jj'} = g\left(\frac{\Delta_{jj'}}{\theta}\right). \tag{3.1}$$

Dans cette dernière expression, θ est appelé le *paramètre de portée* ; celui-ci contrôle la vitesse à laquelle le lien de dépendance fléchit à mesure que la distance entre deux sites devient grande. Le rôle de ce paramètre a été étudié en détails par [9]. Il a été établi par [4] que pour que la matrice Γ_θ soit définie positive pour chaque entier d, il faut que la fonction de lien g respecte la condition que pour tout $\mathrm{s}_1, \ldots, \mathrm{s}_d \in S$ et tout nombres réels $r_1, \ldots, r_d$,

$$\sum_{j,j'=1}^{d} r_j \, r_{j'} \, g(\Delta_{jj'}) > 0.$$

Plusieurs fonctions de lien satisfont ces conditions. La plus populaire est probablement la fonction Matérn définie par

$$g_\nu^{\mathrm{Mat}}(x) = \frac{x^\nu \mathcal{K}_\nu(x)}{2^{\nu-1}\Gamma(\nu)},$$

où $\mathcal{K}_\nu(x)$ est la fonction modifiée de Bessel de second type (voir [24]) et $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{d}t$ est la fonction Gamma. Lorsque $\nu = 1/2$, la fonction Matérn correspond à la fonction de lien Exponentielle, c'est-à-dire $g^{\mathrm{Exp}}(x) = e^{-x}$. Dans ce mémoire, on va

également considérer la fonction de lien *Rationnelle quadratique* $g_\nu^{\mathrm{RQ}}(x) = (1+x^2)^{-\nu}$. Dans les deux cas, $\nu > 0$ joue un rôle au niveau de la souplesse de la fonction.

À partir d'une fonction de lien g donnée, on peut introduire un paramètre d'*effet de pépite*. Ce dernier permet de tenir compte des variations à petite échelle ou encore des erreurs de mesure. Pour ce faire, on définit $\epsilon \in [0,1)$ comme le paramètre de pépite et on travaille avec la fonction de lien modifiée

$$g_\epsilon(x) = \epsilon\, \mathbb{I}\,(x = 0) + (1-\epsilon) g(x). \tag{3.2}$$

De cette façon, lorsque $\epsilon > 0$, la fonction g_ϵ présente un saut à $x = 0$. En effet, d'une part on a $g_\epsilon(0) = \epsilon + (1-\epsilon)g(0) = \epsilon + 1 - \epsilon = 1$, alors que pour $x > 0$ arbitrairement petit, $g_\epsilon(x) = (1-\epsilon)g(x) \leq (1-\epsilon)g(0) = 1-\epsilon$.

3.2.3 Copules spatiales

Le premier auteur à proposer l'utilisation des copules pour faire de la modélisation spatiale est apparamment [1]. Quelques autres travaux ont également adopté cette approche, dont notamment [2], [11], [6] et [18]. Le gain en popularité de cette technique de modélisation spatiale s'explique par le fait que dans le cadre spatial, les copules offrent deux avantages de taille :

(i) elles fournissent des méthodes alternatives à celles basées sur le variogramme ;

(ii) elles permettent la construction de modèles flexibles qui ciblent la dépendance spatiale et qui permettent un grand choix de distributions marginales.

La modélisation de données spatiales à l'aide de copules consiste à caractériser la distribution conjointe $H_{\mathbf{X}}$ du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$. En invoquant le Théorème de Sklar, on sait qu'il existe une copule $C : [0,1]^d \to [0,1]$ telle que

$$H_{\mathbf{X}}(x_1, \ldots, x_d) = C\left\{F_1(x_1), \ldots, F_d(x_d)\right\}.$$

Ainsi, la copule C va caractériser entièrement la dépendance entre les d sites. Ce ne sont toutefois pas tous les modèles de copules qui sont convenables pour la modélisation et l'interpolation spatiales. En fait, tel que noté par [18], deux éléments sont essentiels pour qu'une copule se qualifie comme copule spatiale. D'abord, il faut que cette copule permette une paramétrisation en terme d'une matrice de corrélation qui gère les niveaux de dépendance entre les paires. De plus, le modèle doit permettre le passage entre une dimension $d-1$ à une dimension d afin de procéder à de l'interpolation. Formellement, une copule spatiale d-dimensionnelle sera indicée par une matrice de corrélation Γ, notée $C_{d,\Gamma}$, où on suppose que pour tout $d \in \mathbb{N}$,

$$C_{d,\Gamma}(u_1, \ldots, u_{d-1}, 1) = C_{d-1,\Gamma^\star}(u_1, \ldots, u_{d-1}),$$

où $\Gamma^\star \in \mathbb{R}^{(d-1)\times(d-1)}$ est la matrice de corrélation formée des $d-1$ premières lignes et colonnes de Γ. On suppose également que la copule $C_{d,\Gamma}$ est invariante sous les permutations au sens où pour une matrice de permutations quelconque $\Lambda \in \mathbb{R}^{d\times d}$, la copule de $\Lambda\mathbf{Z}$ est $C_{d,\Lambda\Gamma\Lambda^\top}$.

3.2.4 Copule Normale spatiale et quelques modèles reliés

La copule Normale satisfait les exigences mentionnées à la sous-section précédente, donc elle est appropriée pour la modélisation spatiale. Pour ce faire, il s'agit simplement de la paramétriser à l'aide d'une matrice de corrélation spatiale $\Gamma_\theta \in \mathbb{R}^{d\times d}$ telle que définie à l'Équation (3.1). La copule Normale spatiale est donc

$$C^{\mathrm{N}}_{d,\Gamma_\theta}(u_1, \ldots, u_d) = \Phi_{\Gamma_\theta}\left\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right\}.$$

Une autre copule spatiale émerge si on considère le maximum de K vecteurs aléatoires indépendants de loi Normale. Pour la décrire, soient des vecteurs aléatoires indépendants $\mathbf{X}_1, \ldots, \mathbf{X}_K$ de loi Normale d-dimensionnelle standard de matrice de corrélation

Γ_θ. On définit alors la copule spatiale max-Normale comme la structure de dépendance du maximum *composante par composante* de $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$. On peut montrer que cette copule s'exprime sous la forme

$$C^{\mathrm{Max}}_{d,K,\Gamma_\theta}(u_1, \ldots, u_d) = \left\{ C^{\mathrm{N}}_{d,\Gamma_\theta}\left(u_1^{1/K}, \ldots, u_d^{1/K}\right) \right\}^K.$$

Évidemment, on retrouve la copule Normale spatiale lorsque $K = 1$.

D'autres modèles de copules spatiales peuvent être obtenus en appliquant des transformations non monotones sur les composantes d'un vecteur $(X_1, \ldots, X_d)$ de loi Normale standard de matrice de corrélation Γ_θ. Autrement dit, pour une certaine fonction $\eta : \mathbb{R} \to \mathbb{R}$, on considère la copule du vecteur $(\eta(X_1), \ldots, \eta(X_d))$. Si η est monotone, on retrouve la copule Normale. Les cas intéressants apparaissent plutôt lorsque η est non-monotone. Par exemple, [13] a considéré pour $\lambda \in \mathbb{R}$ la transformation $\eta(x) = x^\lambda \mathrm{I}(x \geq 0) - x\,\mathrm{I}(x < 0)$, ce qui mène à la copule V-transformée. Quand $d = 2$, cette copule est de la forme

$$\begin{aligned} C^{\mathrm{V}}_{\rho,\lambda}(u, v) = {} & \Phi_\rho\left\{\zeta(u)^{1/\lambda}, \zeta(v)^{1/\lambda}\right\} - \Phi_\rho\left\{-\zeta(u), \zeta(v)^{1/\lambda}\right\} \\ & - \Phi_\rho\left\{\zeta(u)^{1/\lambda}, -\zeta(v)\right\} + \Phi_\rho\left\{-\zeta(u), -\zeta(v)\right\}, \end{aligned}$$

où ζ est l'inverse de $\Phi(y^{1/\lambda}) + \Phi(y) - 1$. À noter que lorsque $\lambda = 1$, la transformation est équivalente à $\eta(x) = |x|$. À cause de l'invariance des copules sous des transformations monotones croissantes, on sait que le vecteur aléatoire $(|X_1|, \ldots, |X_d|)$ possède la même copule que $(X_1^2, \ldots, X_d^2)$. Donc, la copule V-transformée dans ce cas particulier n'est nulle autre que la copule Khi-deux centrée. D'ailleurs, plus généralement, les modèles de la famille des copules Khi-deux d-dimensionnelles avec paramètre de décentralité a possèdent les caractéristiques nécessaires à la modélisation spatiale.

3.2.5 Portée efficace

La *portée efficace* $\mathcal{D}$ associée à un modèle donné est définie comme la distance entre deux lieux telle que la valeur du tau de Kendall atteint la valeur $1/5$. Plus formellement, soit une certaine copule spatiale $C_{d,\Gamma_\theta,\theta}$ dont la copule quand $d = 2$ est notée $C_{g(\mathcal{D}/\theta)}$. Alors la portée efficace $\mathcal{D}$ est telle que $\tau(C_{g(\mathcal{D}/\theta)}) = 1/5$. La portée efficace peut être interprétée indépendamment du modèle de copule. En effet, on peut affirmer que la valeur du tau de Kendall pour deux sites séparés par une distance inférieure (resp. supérieure) à la portée efficace $\mathcal{D}$ sera plus élevée (resp. inférieur) que $1/5$. Pour la copule Normale couplée avec une fonction de lien g, la portée efficace $\mathcal{D}$ va satisfaire

$$\frac{2}{\pi}\sin^{-1}\left\{g\left(\frac{\mathcal{D}}{\theta}\right)\right\} = \frac{1}{5},$$

ce qui permet de déduire que

$$\mathcal{D} = \theta\, g^{-1}\left\{\sin\left(\frac{\pi}{10}\right)\right\}.$$

Pour la copule Khi-deux centrée, $\mathcal{D}$ satisfait plutôt

$$\left[\frac{2}{\pi}\sin^{-1}\left\{g\left(\frac{\mathcal{D}}{\theta}\right)\right\}\right]^2 = \frac{1}{5}.$$

Quelques calculs permettent alors de conclure que

$$\mathcal{D} = \theta\, g^{-1}\left\{\sin\left(\frac{\pi}{2\sqrt{5}}\right)\right\}.$$

Plus généralement, en se basant sur l'Équation (2.6), la valeur de $\mathcal{D}$ pour la copule $C^{\chi}_{\rho,a}$ est telle que pour $\rho = g(\mathcal{D}/\theta)$,

$$\left(\frac{2}{\pi}\sin^{-1}\rho\right)\left\{4\Phi_\rho(\sqrt{2}a, \sqrt{2}a) - 4\Phi(\sqrt{2}a) + 1\right\} = \frac{1}{5}.$$

Sauf pour les cas extrêmes $a = 0$ et $a \to \infty$, la solution $\rho^\star$ de cette équation doit s'obtenir numériquement ; on pose ensuite $\mathcal{D} = \theta\, g^{-1}(\rho^\star)$.

3.2.6 Estimation générale des paramètres

Considérons une famille de copules spatiales C_{β,Γ_θ} répondant aux conditions établies au début de cette section. Ici, β est un paramètre spécifique à la famille de copules choisie ; par exemple, $\beta = a$ est le paramètre de décentralité dans le cas de la copule Khi-deux. Les éléments de la matrice de corrélation Γ_θ sont définies implicitement à l'aide d'une fonction de lien g, de sorte que $(\Gamma_\theta)_{jj'} = g_\kappa(\Delta_{jj'}/\theta)$, où $\kappa = (\nu, \epsilon) \in A = (0, \infty) \times [0, 1)$ est un vecteur composé du paramètre de souplesse ν et de l'effet de pépite ϵ, et θ est le paramètre de portée.

Une étape importante consiste à estimer le vecteur des paramètres (θ, κ, β). Pour ce faire, on dispose d'une seule réplique du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$. Dans un tel contexte, il est impossible d'estimer les marges $F_1, \ldots, F_d$, à moins que l'on suppose que le champ aléatoire est stationnaire au sens où $F_1 = \cdots = F_d = F$. Dans ce cas, en posant $C_{\beta,\Gamma_\theta} = C_{\theta,\kappa,\beta}$, la loi conjointe de $\mathbf{X}$ s'écrit sous la forme

$$H_{\theta,\kappa,\beta}(x_1, \ldots, x_d) = C_{\theta,\kappa,\beta}\left\{F(x_1), \ldots, F(x_d)\right\}.$$

La densité associée à $H_{\theta,\kappa,\beta}$ est donc

$$h_{\theta,\kappa,\beta}(x_1, \ldots, x_d) = c_{\theta,\kappa,\beta}\left\{F(x_1), \ldots, F(x_d)\right\} \prod_{j=1}^{d} f(x_j),$$

où $c_{\theta,\kappa,\beta}$ est la densité de $C_{\theta,\kappa,\beta}$ et $f(x) = F'(x)$ est la densité marginale. La stratégie qui est proposée est basée sur une fonction de vraisemblance par paires calculée à partir des densités bivariées. Cette façon de faire évite de considérer la loi complète d-dimensionnelle, qui est souvent difficile à calculer, comme dans le cas de la copule

Khi-deux, qui comporte 2^d termes. Pour décrire la méthode, on définit $C_{\beta,g_\kappa(\Delta/\theta)}$ comme la copule d'une paire de sites tirée du vecteur $\mathbf{X}$ dont la distance les séparant est Δ. La fonction de log-vraisemblance par paires telle que définie par [18] est

$$\mathcal{L}(\theta, \kappa, \beta) = \sum_{1 \leq j < j' \leq d} \ln c_{\beta, g_\kappa(\Delta_{jj'}/\theta)} \left\{ \widehat{F}(X_j), \widehat{F}(X_{j'}) \right\},$$

où $\widehat{F}$ est la fonction de répartition empirique, c'est-à-dire

$$\widehat{F}(x) = \frac{1}{n+1} \sum_{j=1}^{d} \mathrm{I}(X_j \leq x).$$

On obtient ainsi l'estimateur de vraisemblance par paires

$$\left(\widehat{\theta}, \widehat{\kappa}, \widehat{\beta}\right) = \underset{(\theta,\kappa,\beta) \in \mathbb{R}^+ \times A \times B}{\operatorname{argmax}} \mathcal{L}(\theta, \kappa, \beta),$$

où A et B sont les ensembles des valeurs possibles des paramètres κ et β, respectivement. Les fonctions de vraisemblance par paires ont été étudiées en détails par [7] ; il s'agit en fait d'un cas particulier d'une approche plus générale basée sur les fonctions de vraisemblance composites considérées entre autres par [23] et [22]. Ces approches sont très utiles dans les cas où la fonction de vraisemblance complète est complexe et difficile à manipuler. Dans le cas de modèles de copules, cette méthode a été employée par [10] dans le cas de marges connues.

Un autre avantage des fonctions de vraisemblance par paires réside dans la possibilité d'exclure des paires de sites jugés trop éloignés au sens où l'information qu'elles procurent est négligeable. Spécifiquement, en posant $\xi \in (0, 1]$ comme un certain percentile de l'ensemble des $d(d-1)/2$ distances $\Delta_{jj'}$, $1 \leq j < j' \leq d$, on définit D_ξ comme la distance maximale à considérer. Autrement dit, les sites éloignés d'une distance supérieure à D_ξ seront ignorés dans l'expression de la vraisemblance par paires.

Cette dernière s'écrit alors

$$\mathcal{L}^{\xi}(\theta, \kappa, \beta) = \sum_{\substack{1 \leq j < j' \leq d \\ \Delta_{jj'} \leq D_{\xi}}} \ln c_{\beta, g_{\kappa}(\Delta_{jj'}/\theta)} \left\{ \widehat{F}(X_j), \widehat{F}(X_{j'}) \right\}. \tag{3.3}$$

L'estimateur correspondant est donc

$$\left(\widehat{\theta}, \widehat{\kappa}, \widehat{\beta}\right)^{\xi} = \underset{(\theta,\kappa,\beta) \in \mathbb{R}^{+} \times A \times B}{\operatorname{argmax}} \mathcal{L}^{\xi}(\theta, \kappa, \beta).$$

Dans le cas où $\xi = 1$, on retrouve bien entendu l'estimateur $(\widehat{\theta}, \widehat{\kappa}, \widehat{\beta})$. Une étude de simulations exhaustive conduite par [18] a montré que l'usage d'une valeur petite de ξ amène de très bons résultats en général.

Dans les sections subséquentes, on relâchera la contrainte d'un champ aléatoire observé une seule fois en considérant des observations faites successivement dans le temps.

3.3 Un modèle spatio-temporel pour le cas i.i.d. avec marges continues

3.3.1 Construction du modèle

On souhaite modéliser un phénomène tel que les observations disponibles sont des vecteurs d-dimensionnels indépendants $\mathbf{X}_1, \ldots, \mathbf{X}_n$, où $\mathbf{X}_t = (X_{t1}, \ldots, X_{td})$ est le vecteur des observations à tous les sites $j \in \{1, \ldots, d\}$ au temps $t \in \{1, \ldots, n\}$. Pour ce faire, on va supposer que la fonction de répartition marginale associée à la variable aléatoire du j-ème site est continue et appartient à une certaine famille paramétrique $\{F_{\gamma}; \gamma \in \Gamma\}$. On va également supposer que la structure de dépendance spatiale est régie par une copule qui appartient à la famille Khi-deux ; autrement dit, pour chaque

$t \in \{1, \ldots, n\}$, la copule de $\mathbf{X}_t$ est $C^{\chi}_{\Sigma_{\theta,\epsilon},a}$, où la matrice de corrélation $\Sigma_{\theta,\epsilon}$ est telle que pour une certaine fonction de lien g, on a pour g_ϵ définie à l'Équation (3.2) que

$$(\Sigma_{\theta,\epsilon})_{jj'} = g_\epsilon\left(\frac{\Delta_{jj'}}{\theta}\right). \tag{3.4}$$

À cette fin, soit un processus stochastique $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ tel que pour chaque $t \in \mathbb{N}$, $\mathbf{Z}_t$ est distribué selon une loi Normale standard d-dimensionnelle. On suppose ici que $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ sont indépendantes et identiquement distribuées (i.i.d.), c'est-à-dire qu'elles sont stationnaires et indépendantes du temps. On définit ensuite un processus $(\mathbf{X}_t)_{t\in\mathbb{N}}$ tel que pour chaque $j \in \{1, \ldots, d\}$ et $t \in \mathbb{N}$,

$$X_{tj} = F^{-1}_{\gamma_j}\left\{G_a\left((Z_{tj}+a)^2\right)\right\},$$

où $G_a(x) = \Phi(\sqrt{x}+a)+\Phi(\sqrt{x}-a)-1$. La construction de la copule Khi-deux permet de déduire que la loi conjointe de $\mathbf{X}_t = (X_{t1}, \ldots, X_{td})$ pour chaque $t \in \mathbb{N}$ est

$$H_{\theta,\epsilon,a}(\mathbf{x}) = \mathbb{P}(\mathbf{X}_t \leq \mathbf{x}) = C^{\chi}_{\Sigma_{\theta,\epsilon},a}\left\{F_{\gamma_1}(x_1), \ldots, F_{\gamma_d}(x_d)\right\}. \tag{3.5}$$

Le processus $(\mathbf{X}_t)_{t\in\mathbb{N}}$ possède donc, tel que souhaité, une structure de dépendance caractérisée par une copule Khi-deux de matrice spatiale $\Sigma_{\theta,\epsilon}$ et de paramètre de décentralité a. De plus, les marges de $\mathbf{X}_t$ sont respectivement $F_{\gamma_1}, \ldots, F_{\gamma_d}$.

3.3.2 Estimation des paramètres

Supposons que l'on observe n répliques indépendantes $\mathbf{X}_1, \ldots, \mathbf{X}_n$ du processus $(\mathbf{X}_t)_{t\in\mathbb{N}}$. Comme ces répliques sont indépendantes, leur fonction de vraisemblance complète s'écrit comme le produit des n vraisemblances individuelles. On note d'abord que

pour chaque $t \in \{1, \ldots, n\}$, la densité de $\mathbf{X}_t$ s'écrit

$$h_{\theta,\epsilon,a}(x_1, \ldots, x_d) = c^{\chi}_{\Sigma_{\theta,\epsilon},a}\left\{F_{\gamma_1}(x_1), \ldots, F_{\gamma_d}(x_d)\right\} \prod_{j=1}^{d} f_{\gamma_j}(x_j),$$

où f_{γ_j} est la densité associée à F_{γ_j}. La fonction de log-vraisemblance est donc

$$\mathcal{L}(\boldsymbol{\gamma}, \theta, \epsilon, a) = \sum_{t=1}^{n} \left\{ \ln c^{\chi}_{\Sigma_{\theta,\epsilon},a}\left\{F_{\gamma_1}(X_{t1}), \ldots, F_{\gamma_d}(X_{td})\right\} + \sum_{j=1}^{d} \ln f_{\gamma_j}(X_{tj}) \right\},$$

où $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)$ est le vecteur des paramètres des lois marginales et $c^{\chi}_{\Sigma_{\theta,\epsilon},a}$ est la densité de la copule Khi-deux décrite à l'Équation (2.8). On peut donc, en principe du moins, définir les estimateurs à maximum de vraisemblance de $(\boldsymbol{\gamma}, \theta, \epsilon, a)$ par

$$\left(\widehat{\boldsymbol{\gamma}}, \widehat{\theta}, \widehat{\epsilon}, \widehat{a}\right) = \underset{\boldsymbol{\gamma},\theta,\epsilon,a}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\gamma}, \theta, \epsilon, a).$$

Cette approche serait toutefois extrêmement fastidieuse et instable numériquement, car il y a un grand nombre de paramètres et la densité $c^{\chi}_{\Sigma_{\theta,\epsilon},a}$ possède 2^d termes.

Plus efficacement, les paramètres marginaux $\gamma_1, \ldots, \gamma_d$ seront d'abord estimés par maximum de vraisemblance, c'est-à-dire que pour chaque $j \in \{1, \ldots, d\}$,

$$\widehat{\gamma}_j = \underset{\gamma \in \Gamma}{\operatorname{argmax}} \sum_{t=1}^{n} \ln f_{\gamma}(X_{tj}).$$

Ensuite, conditionnellement à ces estimations $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_d$, on maximisera la vraisemblance par paires afin d'estimer (θ, ϵ, a), c'est-à-dire

$$\left(\widehat{\theta}, \widehat{\epsilon}, \widehat{a}\right) = \underset{\theta,\epsilon,a}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{1 \le j < j' \le d} \ln c^{\chi}_{g_\epsilon(\Delta_{jj'}/\theta),a}\left\{F_{\widehat{\gamma}_j}(X_{tj}), F_{\widehat{\gamma}_{j'}}(X_{tj'})\right\}. \tag{3.6}$$

Cependant, puisque les problèmes pratiques qui sont généralement considérés en statistique spatiale comportent un grand nombre de sites, les temps de calcul pour obtenir les estimateurs en (3.6) sont souvent élevés, de l'ordre de plusieurs heures. Une façon

de réduire ce temps de calcul, et aussi dans une certaine mesure, d'améliorer la performance des estimateurs, est de procéder comme à l'Équation (3.3). Ainsi, on pose D_ξ comme la distance maximale à considérer pour un certain percentile $\xi \in (0,1)$ de toutes les distances, et on pose

$$\left(\widehat{\theta}, \widehat{\epsilon}, \widehat{a}\right)^{\xi} = \underset{\theta,\epsilon,a}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{\substack{1 \le j < j' \le d \\ \Delta_{jj'} \le D_\xi}} \ln c^{\chi}_{g_r(\Delta_{jj'}/\theta),a} \left\{ F_{\widehat{\gamma}_j}(X_{tj}), F_{\widehat{\gamma}_{j'}}(X_{tj'}) \right\}. \tag{3.7}$$

3.3.3 Étude de la performance des estimateurs par simulations

L'objectif de cette sous-section est d'étudier la précision de l'estimateur proposé à l'Équation (3.6) sous divers scénarios. Pour ce faire, on va supposer dans la suite que le paramètre a de décentralité est connu ; en fait, on prendra $a \in \{0, 1, \infty\}$. On rappelle que $a = 0$ correspond à la copule Khi-deux centrée, alors que $a \to \infty$ fait apparaître la copule Normale. On va aussi supposer un effet de pépite nul ($\epsilon = 0$), c'est-à-dire qu'il n'y a pas de variations à petite échelle ou d'erreurs de mesure. Ainsi, une fois que les paramètres des marges $\gamma_1, \ldots, \gamma_d$ sont estimés, il reste à estimer le paramètre de portée θ. Comme cas particulier de l'Équation (3.6), on a

$$\widehat{\theta} = \underset{\theta>0}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{1 \le j < j' \le d} \ln c^{\chi}_{g(\Delta_{jj'}/\theta),a} \left\{ F_{\widehat{\gamma}_j}(X_{tj}), F_{\widehat{\gamma}_{j'}}(X_{tj'}) \right\}.$$

Le Tableau 3.1 rapporte les résultats d'une étude de simulation lorsque les fonctions de lien sont Matérn et Rationnelle quadratique avec $\nu = .5$ et $d \in \{2, 3\}$. On a supposé des marges Exponentielles de moyenne 1 ; dans ce cas, les estimateurs à maximum de vraisemblance de $\gamma_1, \ldots, \gamma_d$ sont les moyennes empiriques de chaque échantillon marginal. Pour chaque scénario considéré (c'est-à-dire pour chaque choix de a, g, θ et d), on a sélectionné aléatoirement des points sur une grille à l'intérieur du carré unité.

Comme indicateurs de la performance de $\widehat{\theta}$, on a considéré l'erreur quadratique moyenne

relative (EQMR) et le biais relatif (BR) définis respectivement par

$$\mathrm{EQMR}(\widehat{\theta}) = \mathrm{E}\left\{\left(\frac{\widehat{\theta}}{\theta} - 1\right)^2\right\} \quad \text{et} \quad \mathrm{BR}(\widehat{\theta}) = \mathrm{E}\left(\frac{\widehat{\theta}}{\theta} - 1\right).$$

Puisque des expressions explicites pour $\mathrm{EQMR}(\widehat{\theta})$ et $\mathrm{BR}(\widehat{\theta})$ ne sont pas disponibles, on les estimera à partir de $N = 1\ 000$ répétitions du champ aléatoire. On obtiendra alors $\widehat{\theta}_1, \ldots, \widehat{\theta}_N$ et de là, $\mathrm{EQMR}(\widehat{\theta})$ et $\mathrm{BR}(\widehat{\theta})$ sont estimés respectivement par

$$\widehat{\mathrm{EQMR}}(\widehat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\frac{\widehat{\theta}_i}{\theta} - 1\right)^2\right\} \quad \text{et} \quad \widehat{\mathrm{BR}}(\widehat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{\theta}_i}{\theta} - 1\right). \tag{3.8}$$

Ces estimations ont été effectuées à l'aide du *Parallel computing toolbox* du logiciel Matlab. Grâce à ce système de calculs parallèles, les résultats sont obtenus en environ dix fois moins de temps, si on compare à un ordinateur standard.

À la lumière des résultats du Tableau 3.1, on peut tirer plusieurs conclusions intéressantes. Premièrement, tel qu'attendu, l'augmentation de la taille n amène de meilleures estimations du paramètre de portée θ, autant en terme d'erreurs quadratiques moyennes relatives que du point de vue des biais relatifs. On voit aussi que l'estimateur performe mieux sous la copule Normale, c'est-à-dire quand $a \to \infty$, si on compare aux copule Khi-deux quand $a = 0$ et $a = 1$; néanmoins, les résultats sont bons pour ces deux copules. On remarque également que l'estimateur est nettement plus performant lorsque la fonction de lien est la Rationnelle Quadratique plutôt que Matérn. De façon générale, l'estimateur est meilleur en terme d'EQMR lorsque $\theta = 2$ comparé à $\theta = 1$; les biais sont sensiblement équivalents, toutefois. À noter enfin que le cas à $d = 3$ sites fournit de meilleures estimations de θ que lorsqu'il n'y a que $d = 2$ sites ; ceci s'explique par le simple fait que plus il y a de sites, plus il y a de l'information sur la dépendance, et par le fait même sur le paramètre de portée.

TABLE 3.1 – Estimations, basées sur 1 000 répliques, de l'erreur quadratique moyenne relative (EQM) et du biais relatif (B) de $\widehat{\theta}$ dans le cas de marges Exponentielles de moyenne $\gamma = 1$ sous le modèle i.i.d. avec marges continues

			$d = 2$				$d = 3$			
			$n = 50$		$n = 100$		$n = 50$		$n = 100$	
a	g	θ	EQM	B	EQM	B	EQM	B	EQM	B
0	$g_{.5}^{\text{Mat}}$	1	.2244	.0654	.0645	.0414	.0890	.0638	.0360	.0167
		2	.1523	.0940	.0671	.0405	.0776	.0363	.0362	.0254
	$g_{.5}^{\text{RQ}}$	1	.0466	.0340	.0278	.0110	.0193	.0173	.0105	.0115
		2	.0271	.0385	.0123	.0182	.0161	.0196	.0090	.0120
1	$g_{.5}^{\text{Mat}}$	1	.1151	.1029	.0419	.0277	.0638	.0400	.0345	.0097
		2	.0986	.0802	.0455	.0580	.0588	.0593	.0300	.0053
	$g_{.5}^{\text{RQ}}$	1	.0385	.0226	.0086	.0151	.0177	.0079	.0075	.0104
		2	.0236	.0286	.0092	.0200	.0164	.0137	.0080	.0158
∞	$g_{.5}^{\text{Mat}}$	1	.0753	.0554	.0388	.0314	.0400	.0442	.0205	.0211
		2	.0664	.0618	.0353	.0348	.0428	.0469	.0170	.0163
	$g_{.5}^{\text{RQ}}$	1	.0172	.0319	.0067	.0097	.0134	.0167	.0049	.0139
		2	.0150	.0285	.0069	.0101	.0104	.0166	.0054	.0103

3.4 Un modèle spatio-temporel pour le cas sériel avec marges continues

3.4.1 Construction du modèle général

On reprend ici le contexte d'un phénomène observé via des vecteurs à d dimensions $\mathbf{X}_1, \ldots, \mathbf{X}_n$, où $\mathbf{X}_t = (X_{t1}, \ldots, X_{td})$ est le vecteur des observations à tous les sites au

temps $t \in \{1, \ldots, n\}$. Toutefois, contrairement au cas i.i.d. considéré à la Section 3.3, ces vecteurs peuvent être dépendants du temps, c'est-à-dire qu'il y a de la dépendance sérielle. Pour construire un modèle approprié pour cette situation, on reprend l'ingrédient de base du modèle spatio-temporel dans le cas i.i.d. en considérant un processus Gaussien $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ tel que pour chaque $t \in \mathbb{N}$, $\mathbf{Z}_t$ est distribué selon une loi Normale standard d-dimensionnelle de matrice de corrélation $\Sigma_{\theta,\epsilon}$ telle que définie à l'Équation (3.4). Bien qu'on ne le spécifie pas de manière explicite pour l'instant, $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$, sont dépendantes, contrairement au cas i.i.d. développé à la Section 3.3. Ensuite, à partir de $(\mathbf{Z}_t)_{t\in\mathbb{N}}$, on définit un processus $(\mathbf{X}_t)_{t\in\mathbb{N}}$ de telle manière que pour chaque $j \in \{1, \ldots, d\}$ et $t \in \mathbb{N}$,

$$X_{tj} = F_{\gamma_j}^{-1}\left\{G_a\left((Z_{tj} + a)^2\right)\right\}.$$

À l'instar du cas i.i.d., le processus stochastique $\mathbf{X}_t$ est tel que pour chaque $t \in \mathbb{N}$, sa loi conjointe est la même que celle décrite à l'Équation (3.5), à savoir $H_{\theta,\epsilon,a}$. La différence fondamentale réside dans le fait qu'ici, $\mathbf{X}_1, \mathbf{X}_2, \ldots$ sont dépendants.

Soient maintenant $\Omega_1, \ldots, \Omega_d \in \mathbb{R}^{n\times n}$, des matrices de corrélation telles que

$$(\Omega_j)_{tt'} = \operatorname{corr}\left(Z_{tj}, Z_{t'j}\right).$$

Puisque le processus $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ est stationnaire, l'élément en position (t, t') de Ω_j va, en fait, ne dépendre que du délai $|t - t'|$. Typiquement, Ω_j va dépendre d'un vecteur de paramètres $\boldsymbol{\beta}_j$, comme dans le cas de processus autorégressifs ou à moyenne mobile ; on écrira alors $\Omega_j = \Omega_j(\boldsymbol{\beta}_j)$. On en déduit donc que pour chaque $j \in \{1, \ldots, d\}$, la distribution conjointe de la série marginale $X_{1j}, \ldots, X_{nj}$ est donnée par

$$H_{\beta_j,\gamma_j,a}(x_1, \ldots, x_n) = C^{\chi}_{\Omega_j(\boldsymbol{\beta}_j),a}\left\{F_{\gamma_j}(x_1), \ldots, F_{\gamma_j}(x_n)\right\}. \tag{3.9}$$

3.4.2 Estimation des paramètres dans le cas général

Pour le modèle général, les paramètres à estimer sont $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ et $\gamma_1, \ldots, \gamma_d$, associés aux séries marginales, de même que le paramètre de portée θ, l'effet de pépite ϵ et le paramètre de décentralité a. Pour les estimer, on procèdera en trois étapes, à savoir

- $(\mathcal{E}_1)$ qu'on estimera d'abord les paramètres marginaux $\gamma_1, \ldots, \gamma_d$;
- $(\mathcal{E}_2)$ ensuite, conditionnellement aux estimations $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_d$ obtenues à l'étape $\mathcal{E}_1$, on procédera à l'estimation des paramètres de dépendance spatiale θ, ϵ et a ;
- $(\mathcal{E}_3)$ enfin, conditionnellement à $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_d$ et à $\widehat{a}$, on estimera pour chaque $j \in \{1, \ldots, d\}$ le vecteur des paramètres de dépendance sérielle $\boldsymbol{\beta}_j$.

Plus spécifiquement, à partir de l'Équation (3.5), on tire que la densité de $\mathbf{X}_t$ est

$$h_{\theta,\epsilon,a}(x_1, \ldots, x_d) = c^{\chi}_{\Sigma_{\theta,\epsilon},a}\left\{F_{\gamma_1}(x_1), \ldots, F_{\gamma_d}(x_d)\right\} \prod_{j=1}^{d} f_{\gamma_j}(x_j).$$

Pour simplifier le problème, on estimera au préalable le paramètre γ_j par maximum de vraisemblance. Ensuite, conditionnellement à ces estimations $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_d$, on définira les estimateurs de θ, ϵ et a en utilisant une fonction de pseudo-vraisemblance composite par paires sous l'hypothèse d'indépendance sérielle, à savoir

$$\left(\widehat{\theta}, \widehat{\epsilon}, \widehat{a}\right) = \underset{\theta,\epsilon,a}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{1 \le j < j' \le d} \ln c_{g_\epsilon(\Delta_{jj'}/\theta),a}\left\{F_{\widehat{\gamma}_j}(X_{tj}), F_{\widehat{\gamma}_{j'}}(X_{tj'})\right\}.$$

Cette pseudo-vraisemblance est en fait la même que celle utilisée dans le cas i.i.d. ; son usage est possible dans le cas d'une dépendance sérielle car les paramètres $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ y sont absents. À noter que l'usage de vraisemblances composites comme celle ci-dessus a fait l'objet de nombreux travaux qui montrent sa validité, sous certaines conditions non-contraignantes ; un excellent survol de ces méthodes se retrouve d'ailleurs dans l'article de [22]. Pour l'estimation de $\boldsymbol{\beta}_j$, on note d'abord que l'Équation (3.9) permet

de déduire que la densité conjointe associée à la série marginale $X_{1j}, \ldots, X_{nj}$ est

$$h_{\boldsymbol{\beta}_j,\gamma_j,a}(x_1,\ldots,x_n) = c^{\chi}_{\Omega_j(\boldsymbol{\beta}_j),a}\left\{F_{\gamma_j}(x_1),\ldots,F_{\gamma_j}(x_n)\right\}\prod_{t=1}^{n} f_{\gamma_j}(x_t).$$

La fonction de log-vraisemblance complète de $X_{1j}, \ldots, X_{nj}$ est donc

$$\mathcal{L}(\boldsymbol{\beta}_j, a, \gamma_j) = \ln c^{\chi}_{\Omega_j(\boldsymbol{\beta}_j),a}\left\{F_{\gamma_j}(X_{1j}),\ldots,F_{\gamma_j}(X_{nj})\right\} + \sum_{t=1}^{n} \ln f_{\gamma_j}(X_{tj}).$$

Puisque les paramètres γ_j et a ont été préalablement estimés aux étapes $\mathcal{E}_1$ et $\mathcal{E}_2$, l'estimation de $\boldsymbol{\beta}_j$ est donnée pour chaque $j \in \{1,\ldots,d\}$ par

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j}{\operatorname{argmax}}\left[\ln c^{\chi}_{\Omega_j(\boldsymbol{\beta}_j),\widehat{a}}\left\{F_{\widehat{\gamma}_j}(X_{1j}),\ldots,F_{\widehat{\gamma}_j}(X_{nj})\right\}\right]. \tag{3.10}$$

3.4.3 Le cas particulier d'un modèle sous-jacent MM(1)

Supposons que le processus Gaussien $\{\mathbf{Z}_t\}_{t\in\mathbb{N}}$ qui sert d'ingrédient de base à notre modèle stochastique est généré par un processus *à moyenne mobile* d'ordre un, noté MM(1). Autrement dit, on a pour $\beta_1, \ldots, \beta_d \in (-1, 1)$ que

$$Z_{tj} = \beta_j\, \varepsilon_{t-1,j} + \varepsilon_{tj},$$

où $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{N}}$ est un processus d'innovations indépendantes tel que pour chaque $t \in \mathbb{N}$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \ldots, \varepsilon_{td})$ suit une loi Normale d-dimensionnelle de moyennes nulles. De plus, afin de satisfaire les exigences du modèle général, on suppose que la matrice de variance-covariance $\Sigma_{\boldsymbol{\varepsilon}}$ de $\boldsymbol{\varepsilon}_t$ est telle que ses éléments diagonaux sont

$$(\Sigma_{\boldsymbol{\varepsilon}})_{jj} = \frac{1}{\beta_j^2 + 1},$$

alors que pour $j \neq j'$,

$$(\Sigma_\epsilon)_{jj'} = \frac{1}{\beta_j \beta_{j'} + 1} \, g_\epsilon \left(\frac{\Delta_{jj'}}{\theta} \right).$$

De cette façon, on a tel qu'exigé que si $j \neq j'$,

$$\begin{aligned}
(\Sigma_{\theta,\epsilon})_{jj'} &= \operatorname{cov}\left(Z_{tj}, Z_{tj'}\right) \\
&= \operatorname{cov}\left(\beta_j \, \varepsilon_{t-1,j} + \varepsilon_{tj}, \beta_{j'} \, \varepsilon_{t-1,j'} + \varepsilon_{tj'}\right) \\
&= \operatorname{cov}\left(\beta_j \, \varepsilon_{t-1,j}, \beta_{j'} \, \varepsilon_{t-1,j'}\right) + \operatorname{cov}\left(\varepsilon_{tj}, \varepsilon_{tj'}\right) \\
&= \beta_j \, \beta_{j'} \operatorname{cov}\left(\varepsilon_{t-1,j}, \varepsilon_{t-1,j'}\right) + \operatorname{cov}\left(\varepsilon_{tj}, \varepsilon_{tj'}\right) \\
&= \beta_j \, \beta_{j'} \operatorname{cov}\left(\varepsilon_{tj}, \varepsilon_{tj'}\right) + \operatorname{cov}\left(\varepsilon_{tj}, \varepsilon_{tj'}\right) \\
&= (\beta_j \beta_{j'} + 1) \times \frac{1}{\beta_j \beta_{j'} + 1} \, g_\epsilon \left(\frac{\Delta_{jj'}}{\theta} \right) \\
&= g_\epsilon \left(\frac{\Delta_{jj'}}{\theta} \right).
\end{aligned}$$

Par un calcul similaire, $(\Sigma_{\theta,\epsilon})_{jj} = 1$, ce qui montre que $\Sigma_{\theta,\epsilon}$ est bel et bien une matrice de corrélation spatiale. De plus, on a pour chaque $j \in \{1, \ldots, d\}$ que

$$\begin{aligned}
\left(\Omega_j(\beta_j)\right)_{tt'} &= \operatorname{corr}\left(Z_{tj}, Z_{t'j}\right) \\
&= \operatorname{corr}\left(\beta_j \, \varepsilon_{t-1,j} + \varepsilon_{tj}, \beta_j \, \varepsilon_{t'-1,j} + \varepsilon_{t',j}\right) \\
&= \begin{cases} 1, & \text{si } t = t'; \\ \dfrac{\beta_j}{\beta_j^2 + 1}, & \text{si } |t - t'| = 1; \\ 0, & \text{sinon.} \end{cases}
\end{aligned}$$

Le modèle présente donc une structure de dépendance sérielle de délai 1. Plus précisément, on a pour chaque $j \in \{1, \ldots, d\}$ que la loi conjointe de $(X_{tj}, X_{t+1,j})$ possède une copule Khi-deux de paramètre $\beta_j/(\beta_j^2 + 1)$ et des marges identiques F_{γ_j}.

3.4.4 Estimation des paramètres pour le modèle MM(1)

Tel que mentionné précédemment, l'estimation des paramètres $\gamma_1, \ldots, \gamma_d$, θ, ϵ et a s'effectue aux étapes $\mathcal{E}_1$ et $\mathcal{E}_2$, indépendamment de la structure de dépendance sérielle du processus. Toutefois, le fait que les processus à moyenne mobile d'ordre un induisent une dépendance de délai un permet de simplifier la fonction de vraisemblance de $X_{1j}, \ldots, X_{nj}$. En effet, on a alors que leur loi conjointe est donnée par

$$h_{\beta_j,\gamma_j,a}(X_{1j}, \ldots, X_{nj}) = \prod_{t=1}^{n-1} c^{\chi}_{\frac{\beta_j}{\beta_j^2+1},a}\left\{F_{\gamma_j}(X_{tj}), F_{\gamma_j}(X_{t+1,j})\right\} \times \prod_{t=1}^{n} f_{\gamma_j}(X_{tj}).$$

Par conséquent, conditionnellement à $\widehat{\gamma}_j$ et $\widehat{a}$,

$$\widehat{\beta}_j = \underset{\beta_j \in (-1,1)}{\operatorname{argmax}} \sum_{t=1}^{n-1} \ln c^{\chi}_{\frac{\beta_j}{\beta_j^2+1},\widehat{a}}\left\{F_{\widehat{\gamma}_j}(X_{tj}), F_{\widehat{\gamma}_j}(X_{t+1,j})\right\}.$$

Le Tableau 3.2 présente les résultats d'une étude de simulations quant à la performance de $\widehat{\theta}$ dans le cas d'une dépendance sérielle de type MM(1). À l'instar du Tableau 3.1, on a fixé a, ainsi que $\epsilon = 0$. Ainsi, cette étude permet de vérifier dans quelle mesure le fait de considérer l'indépendance sérielle pour l'estimation de θ a une influence sur la précision de $\widehat{\theta}$. Ici, on a considéré seulement le cas $d = 2$ et $n = 50$, mais comme pour le Tableau 3.1, les résultats peuvent être transférés à des dimensions d et à des tailles d'échantillons n supérieures. À noter aussi que le cas $a = 1$ a été ignoré dans cette simulation car l'obtention des résultats est extrêmement longue, toujours de l'ordre de plusieurs heures. Quelques expériences nous ont toutefois convaincu que l'estimateur $\widehat{\theta}$ se comporte bien en termes d'erreur quadratique moyenne et de biais dans ce cas.

De façon générale, on peut émettre des commentaires similaires à ceux concernant les résultats du Tableau 3.1 sur l'influence de la fonction de lien et la valeur prise par θ. Cependant, on remarque que les erreurs quadratiques moyennes relatives et les biais relatifs sont généralement plus élevés ici dans le cas sériel que pour le cas i.i.d., donc

TABLE 3.2 – Estimations, basées sur 1 000 répliques, de l'erreur quadratique moyenne relative (EQM) et du biais relatif (B) de $\widehat{\theta}$ dans le cas de marges Exponentielles de moyenne $\gamma = 1$ sous un processus sériel MM(1) quand $d = 2$; panneau supérieur : $n = 50$; panneau inférieur : $n = 100$

			$\beta_1 = \beta_2 = .25$		$\beta_1 = \beta_2 = .50$		$\beta_1 = \beta_2 = .75$	
a	g	θ	EQM	B	EQM	B	EQM	B
0	$g_{.5}^{\text{Mat}}$	1	.1958	.0926	.4638	.0829	.3097	.1509
		2	.1410	.0850	.2151	.1155	.2407	.1207
	$g_{.5}^{\text{RQ}}$	1	.0246	.0245	.0470	.0377	.0438	.0398
		2	.0272	.0246	.0271	.0227	.0333	.0275
∞	$g_{.5}^{\text{Mat}}$	1	.0779	.0751	.1730	.1009	.1126	.0826
		2	.0927	.0670	.1220	.1013	.1456	.1187
	$g_{.5}^{\text{RQ}}$	1	.0216	.0374	.0264	.0444	.0247	.0397
		2	.0174	.0278	.0237	.0427	.0242	.0402
0	$g_{.5}^{\text{Mat}}$	1	.0638	.0429	.2686	-.0047	.0860	.0617
		2	.0715	.0345	.0658	.0613	.0727	.0545
	$g_{.5}^{\text{RQ}}$	1	.0166	.0177	.0327	.0331	.0191	.0199
		2	.0141	.0132	.0148	.0149	.0180	.0181
∞	$g_{.5}^{\text{Mat}}$	1	.0641	.0367	.0465	.0428	.0943	.0489
		2	.0453	.0594	.0532	.0640	.0555	.0418
	$g_{.5}^{\text{RQ}}$	1	.0100	.0110	.0106	.0180	.0136	.0277
		2	.0082	.0116	.0096	.0182	.0107	.0196

quand $\beta_1 = \beta_2 = 0$. En fait, on voit que l'imprécision de $\widehat{\theta}$ augmente à mesure que le niveau de dépendance sérielle, contrôlé par β_1, β_2, augmente. Ce comportement s'explique possiblement par le fait que plus β_1, β_2 sont élevés, plus la vraisemblance composite qui est utilisée s'éloigne de la véritable fonction de vraisemblance. À noter

finalement que la performance de l'estimateur s'améliore sensiblement lorsque la taille des échantillons augmente.

Une étude de simulations a aussi été conduite concernant l'efficacité de l'estimateur du paramètre sériel $\widehat{\beta}$; les résultats se retrouvent dans le Tableau 3.3. De façon générale, l'estimateur se comporte très bien, surtout au niveau du biais relatif. On constate aussi que ses performances s'améliorent lorsque la taille n augmente. Il est également meilleur sous une copule Normale ($a \to \infty$) que sous une copule Khi-deux centrée.

TABLE 3.3 – Estimations, basées sur 1 000 répliques, de l'erreur quadratique moyenne relative (EQM) et du biais relatif (B) de $\widehat{\beta}$ dans le cas de marges Exponentielles de moyenne $\gamma = 1$ sous un processus sériel MM(1)

		$n = 50$		$n = 100$	
a	β	EQM	B	EQM	B
0	0,2	2.465	0.161	1.771	0.125
	0,4	0.899	0.031	0.592	0.000
	0,6	0.417	-0.072	0.289	-0.013
∞	0,2	0.747	0.035	0.309	0.020
	0,4	0.358	0.096	0.161	0.041
	0,6	0.191	0.065	0.141	0.084

3.5 Un modèle spatio-temporel pour le cas i.i.d. avec marges de pluie

3.5.1 Construction du modèle

Jusqu'ici, on a supposé des marges continues. Toutefois, il arrive que certains phénomènes naturels conduisent à des situations où une ou plusieurs valeurs apparaissent avec une masse de probabilité non-nulle. Une telle situation se produit pour des données de précipitation, où le nombre de zéros observés, qui correspondent à une absence de pluie, est non négligeable. Afin d'en tenir compte, il faudra adapter à la fois notre modèle ainsi que la procédure d'estimation des paramètres.

Supposons donc que la densité de la loi à un site $j \in \{1, \ldots, d\}$ est de la forme mixte

$$f_{p_j,\gamma_j}(x) = p_j^{\mathbb{I}(x=0)} \left\{(1-p_j) f_{\gamma_j}(x)\right\}^{\mathbb{I}(x>0)},$$

où f_{γ_j} est une densité de probabilité définie sur $(0,\infty)$ qui correspond à la partie continue et le paramètre $p_j \in [0,1]$ représente la probabilité d'observer un zéro. On montre que la fonction de répartition associée à cette densité est

$$F_{p_j,\gamma_j}(x) = p_j + (1-p_j) F_{\gamma_j}(x),$$

où F_{γ_j} est la fonction de répartition associée à f_{γ_j}. On peut montrer que

$$F^{-1}_{p_j,\gamma_j}(u) = \begin{cases} 0, & \text{si } u \in [0, p_j]; \\ F^{-1}_{\gamma_j}\left(\dfrac{u-p_j}{1-p_j}\right), & \text{si } u \in (p_j, 1]. \end{cases}$$

Lemme 3.1. *Pour tout $u \in [0,1]$ et $x \geq 0$, $F^{-1}_{p,\gamma}(u) \leq x$ si et seulement si $u \leq F_{p,\gamma}(x)$.*

Une conséquence directe du Lemme 3.1 est que si $(U_1, \ldots, U_d) \sim C$ et qu'on pose, pour chaque $j \in \{1, \ldots, d\}$,

$$X_j = F^{-1}_{p_j,\gamma_j}(U_j) = F^{-1}_{\gamma_j}\left\{\frac{\max(U_j - p_j, 0)}{1 - p_j}\right\},$$

alors la fonction de répartition conjointe de $(X_1, \ldots, X_d)$ est

$$H(x_1, \ldots, x_d) = C\left\{F_{p_1,\gamma_1}(x_1), \ldots, F_{p_d,\gamma_d}(x_d)\right\}.$$

En effet, par un calcul direct,

$$\begin{aligned} H(x_1, \ldots, x_d) &= \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) \\ &= \mathbb{P}\left\{F^{-1}_{p_1,\gamma_1}(U_1) \leq x_1, \ldots, F^{-1}_{p_d,\gamma_d}(U_j) \leq x_d\right\} \\ &= \mathbb{P}\left\{U_1 \leq F^{-1}_{p_1,\gamma_1}(x_1), \ldots, U_d \leq F^{-1}_{p_d,\gamma_d}(x_j)\right\} \\ &= C\left\{F_{p_1,\gamma_1}(x_1), \ldots, F_{p_d,\gamma_d}(x_d)\right\}. \end{aligned}$$

Ainsi, on peut facilement simuler des données spatiales à partir d'un champ aléatoire de copule Khi-deux et de marges de pluie $F_{p_1,\gamma_1}, \ldots, F_{p_d,\gamma_d}$. La différence fondamentale ici comparée au cas de marges continues est que la densité associée à H comporte des termes supplémentaires dus à la composante discrète des marges. En fait, l'expression de la densité de H dans le cas général est plutôt compliquée. Toutefois, quand $d = 2$, on montre que pour $u_1 = F_{p_1,\gamma_1}(x_1)$ et $u_2 = F_{p_2,\gamma_2}(x_2)$, la densité est donnée par

$$\begin{aligned} h(x_1, x_2) &= \{C(p_1, p_2)\}^{\mathbb{I}(x_1=0, x_2=0)} \\ &\quad \times \left\{C^{01}(p_1, u_2) f_{p_2,\gamma_2}(x_2)\right\}^{\mathbb{I}(x_1=0, x_2>0)} \\ &\quad \times \left\{C^{10}(u_1, p_2) f_{p_1,\gamma_1}(x_1)\right\}^{\mathbb{I}(x_1>0, x_2=0)} \\ &\quad \times \left\{c(u_1, u_2) f_{p_1,\gamma_1}(x_1) f_{p_2,\gamma_2}(x_2)\right\}^{\mathbb{I}(x_1>0, x_2>0)}, \end{aligned}$$

où c est la densité de la copule C et $C^{10}(u_1, u_2) = \partial C(u_1, u_2)/\partial u_1$, $C^{01}(u_1, u_2) = \partial C(u_1, u_2)/\partial u_2$ sont ses dérivées partielles.

3.5.2 Estimation des paramètres

Par des arguments similaires au cas de marges continues, on procèdera d'abord à l'estimation des paramètres marginaux. Ici, pour chaque site $j \in \{1, \ldots, d\}$, il faudra estimer (p_j, γ_j). L'estimateur à maximum de vraisemblance est donné par

$$(\widehat{p}_j, \widehat{\gamma}_j) = \underset{p \in (0,1), \gamma \in \Gamma}{\operatorname{argmax}} \sum_{t=1}^{n} \ln p_j^{\mathbb{I}(X_{tj}=0)} \left\{(1-p_j) f_{\gamma_j}(X_{tj})\right\}^{\mathbb{I}(X_{tj}>0)}. \tag{3.11}$$

À ce titre, le résultat suivant s'avérera très utile.

Proposition 3.1. *L'estimateur à maximum de vraisemblance* $(\widehat{p}_j, \widehat{\gamma}_j)$ *en* (3.11) *est tel que* $\widehat{p}_j$ *est la proportion de zéros dans l'échantillon* $X_{1j}, \ldots, X_{nj}$*, c'est-à-dire*

$$\widehat{p}_j = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}(X_{tj} = 0),$$

alors que $\widehat{\gamma}_j$ *est l'estimateur à maximum de vraisemblance basé sur les* $X_{tj} > 0$.

Maintenant, conditionnellement aux estimations $(\widehat{p}_1, \widehat{\gamma}_1), \ldots, (\widehat{p}_d, \widehat{\gamma}_d)$, on maximisera la vraisemblance par paires afin d'estimer (θ, ϵ, a). Pour ce faire, notons d'abord que pour chaque $t \in \mathbb{N}$, la densité $h_{\theta,\epsilon,a}$ de $(X_{tj}, X_{tj'})$ lorsque $C = C^{\chi}_{g_\epsilon(\Delta_{jj'}/\theta),a}$ est telle que

pour $u_j = F_{p_j,\gamma_j}(x_j)$ et $u_{j'} = F_{p_{j'},\gamma_{j'}}(x_{j'})$,

$$\begin{aligned}
\ln h_{\theta,\epsilon,a}(x_j, x_{j'}) &= \mathbb{I}(x_j = 0, x_{j'} = 0) \ln C^{\chi}_{g_\epsilon(\Delta_{jj'}/\theta),a}(p_j, p_{j'}) \\
&\quad + \mathbb{I}(x_j = 0, x_{j'} > 0) \ln \left\{ C^{\chi,01}_{g_\epsilon(\Delta_{jj'}/\theta),a}(p_j, u_{j'}) f_{p_{j'},\gamma_{j'}}(x_{j'}) \right\} \\
&\quad + \mathbb{I}(x_j > 0, x_{j'} = 0) \ln \left\{ C^{\chi,10}_{g_\epsilon(\Delta_{jj'}/\theta),a}(u_j, p_{j'}) f_{p_j,\gamma_1}(x_j) \right\} \\
&\quad + \mathbb{I}(x_j > 0, x_{j'} > 0) \ln \left\{ c^{\chi}_{g_\epsilon(\Delta_{jj'}/\theta),a}(u_j, u_{j'}) f_{p_j,\gamma_j}(x_j) f_{p_{j'},\gamma_{j'}}(x_{j'}) \right\} \\
&= \mathbb{I}(x_j = 0, x_{j'} = 0) \ln C^{\chi}_{g_\epsilon(\Delta_{jj'}/\theta),a}(p_j, p_{j'}) \\
&\quad + \mathbb{I}(x_j = 0, x_{j'} > 0) \ln C^{\chi,01}_{g_\epsilon(\Delta_{jj'}/\theta),a}(p_j, u_{j'}) \\
&\quad + \mathbb{I}(x_j > 0, x_{j'} = 0) \ln C^{\chi,10}_{g_\epsilon(\Delta_{jj'}/\theta),a}(u_j, p_{j'}) \\
&\quad + \mathbb{I}(x_j > 0, x_{j'} > 0) \ln c^{\chi}_{g_\epsilon(\Delta_{jj'}/\theta),a}(u_j, u_{j'}) + R \\
&= \ln \widetilde{h}_{\theta,\epsilon,a}(x_j, x_{j'}) + R,
\end{aligned}$$

où R ne contient que des termes en $(p_1, \gamma_1), \ldots, (p_d, \gamma_d)$. Comme ces paramètres ont été préalablement estimés, on peut alors négliger R dans l'élaboration de la vraisemblance par paires. On a ainsi

$$\left(\widehat{\theta}, \widehat{\epsilon}, \widehat{a}\right) = \underset{\theta,\epsilon,a}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{1 \le j < j' \le d} \ln \widetilde{h}_{\theta,\epsilon,a}(X_{tj}, X_{tj'}). \tag{3.12}$$

À noter que de l'Équation (2.2), on déduit

$$\begin{aligned}
C^{\chi,10}_{\rho,a}(u_1, u_2) &= \frac{\partial}{\partial u_1} C^{\chi}_{\rho,a}(u_1, u_2) \\
&= \sum_{(\epsilon_1,\epsilon_2) \in \{-1,1\}^2} \epsilon_2 \, C^{\mathrm{N},10}_{\rho} \left\{ \widetilde{h}_a(\epsilon_1 u_1), \widetilde{h}_a(\epsilon_2 u_2) \right\} \widetilde{h}'_a(\epsilon_1 u_1) \\
&= \widetilde{h}'_a(u_1) \left\{ C^{\mathrm{N},10}_{\rho}\left(\widetilde{h}_a(u_1), \widetilde{h}_a(u_2)\right) - C^{\mathrm{N},10}_{\rho}\left(\widetilde{h}_a(u_1), \widetilde{h}_a(-u_2)\right) \right\} \\
&\quad + \widetilde{h}'_a(-u_1) \left\{ C^{\mathrm{N},10}_{\rho}\left(\widetilde{h}_a(-u_1), \widetilde{h}_a(u_2)\right) - C^{\mathrm{N},10}_{\rho}\left(\widetilde{h}_a(-u_1), \widetilde{h}_a(-u_2)\right) \right\},
\end{aligned}$$

où $C^{\mathrm{N},10}_{\rho}(u_1, u_2) = \partial C^{\mathrm{N}}_{\rho}(u_1, u_2)/\partial u_1$. Comme $C^{\chi}_{\rho,a}$ est symétrique, on a

$$C^{\chi,01}_{\rho,a}(u_1, u_2) = \frac{\partial}{\partial u_2} C^{\chi}_{\rho,a}(u_1, u_2) = \frac{\partial}{\partial u_2} C^{\chi}_{\rho,a}(u_2, u_1) = C^{\chi,10}_{\rho,a}(u_2, u_1).$$

Enfin, à noter que $\widetilde{h}'_a(u) = \partial\,\Phi \circ h_a(u)/\partial u = \phi \circ h_a(u) \times h'_a(u)$.

3.5.3 Étude de la performance des estimateurs par simulations

On va conduire ici une étude de simulations similaire à celle du Tableau 3.1, sauf qu'on va considérer des marges discontinues de telle sorte qu'il y a une masse de probabilité non-nulle en zéro. Ainsi, on va prendre des marges Exponentielles pour la partie continue, mais avec $p_1, \ldots, p_d \in \{.2, .4\}$. Donc, il faut d'abord estimer (γ_j, p_j) pour chaque $j \in \{1, \ldots, d\}$, et ensuite, conditionnellement à ces estimations, on obtient comme cas particulier de l'Équation (3.12) avec a fixé et $\epsilon = 0$ que

$$\widehat{\theta} = \underset{\theta>0}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{1\le j<j'\le d} \ln \widetilde{h}_{\theta,0,a}(X_{tj}, X_{tj'}).$$

Les résultats obtenus concernant $\mathrm{EQMR}(\widehat{\theta})$ et $\mathrm{BR}(\widehat{\theta})$ définis à l'Équation (3.8) se retrouvent dans le Tableau 3.4 pour $d \in \{2, 3\}$ et $a \in \{0, \infty\}$.

Les résultats du Tableau 3.4 sont en plusieurs points semblables à ceux du Tableau 3.1. En effet, plus la taille échantillonnale n est élevée, meilleure est la précision de l'estimateur $\widehat{\theta}$. Aussi, la fonction de lien Rationnelle Quadratique est associée à de meilleures précisions en terme d'erreur quadratique moyenne relative comparativement à la fonction Matérn. Tout comme dans le cas i.i.d. avec marges continues, un nombre plus élevé de sites correspond à de meilleurs résultats. Enfin, on remarque que l'estimateur de θ est généralement plus précis quand $p = .2$ comparé à $p = .4$.

TABLE 3.4 – Estimations, basées sur 1 000 répliques, de l'erreur quadratique moyenne relative (EQM) et du biais relatif (B) de $\widehat{\theta}$ sous le modèle i.i.d. avec marges de pluie Exponentielles de moyenne $\gamma = 1$ et de probabilité $p \in \{.2, .4\}$

			$d = 2$				$d = 3$			
			$n = 50$		$n = 100$		$n = 50$		$n = 100$	
a	g	p	EQM	B	EQM	B	EQM	B	EQM	B
0	$g_{.5}^{\text{Mat}}$	.2	.1853	.0955	.1018	.0355	.0958	.0538	.0368	.0275
		.4	.3582	.0276	.0765	.0484	.1046	.0414	.0492	.0254
	$g_{.5}^{\text{RQ}}$	.2	.0286	.0208	.0183	.0147	.0211	.0081	.0127	.0116
		.4	.0722	.0136	.0240	.0186	.0295	.0224	.0153	.0121
∞	$g_{.5}^{\text{Mat}}$	.2	.1335	.0783	.0584	.0261	.0686	.0345	.0316	.0261
		.4	.1634	.0533	.0765	.0519	.0964	.0603	.0369	.0146
	$g_{.5}^{\text{RQ}}$	.2	.0382	.0311	.0215	.0154	.0226	.0182	.0121	.0068
		.4	.0481	.0425	.0240	.0192	.0318	.0213	.0154	.0094

3.6 Un modèle spatio-temporel pour le cas sériel avec marges de pluie

Avec les outils développés jusqu'ici, il est facile de construire un modèle spatio-temporel qui comporte à la fois une composante temporelle et des marges de pluie. Celui-ci sera donc bien adapté pour traiter des phénomènes de précipitations. Pour ce faire, il s'agit de suivre la même logique que celle employée pour la construction du modèle général de la Section 3.4, mais en utilisant des marges de pluie $F_{p_1,\gamma_1}, \ldots, F_{p_d,\gamma_d}$ plutôt que des marges continues $F_{\gamma_1}, \ldots, F_{\gamma_d}$.

Pour l'estimation des paramètres $(p_1, \gamma_1), \ldots, (p_d, \gamma_d)$, on procède de la même façon que pour le cas i.i.d. tel que décrite à la Section 3.5. On procède également de façon identique à celle de la Section 3.5 pour l'estimation des paramètres de dépendance

spatiale θ, ϵ et a, puisqu'on adoptera l'approche composite qui consiste à supposer l'indépendance sérielle. Une fois ces étapes accomplies, il reste donc à estimer les paramètres $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ associés à la dépendance sérielle. Ainsi, conditionnellement aux estimations $(\widehat{p}_1, \widehat{\gamma}_1), \ldots, (\widehat{p}_d, \widehat{\gamma}_d)$ et $\widehat{a}$, on estime individuellement les paramètres $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ en utilisant une version de l'estimateur de l'Équation (3.10) adapté aux marges de pluie. Dans ce cas, le fait de considérer les parties discrètes et continues des marges force à remplacer la densité de copule $c^{\chi}_{\Omega_j(\boldsymbol{\beta}_j),\widehat{a}}$ par une densité mixte qui comporte 2^n termes. Dans le cas du modèle basé sur un processus MM(1), on aura

$$\widehat{\beta}_j = \underset{\beta_j \in (-1,1)}{\operatorname{argmax}} \sum_{t=1}^{n-1} \ln h_{\beta_j}(X_{tj}, X_{t+1,j}),$$

où pour $u = F_{\widehat{p}_j,\widehat{\gamma}_j}(x)$ et $v = F_{\widehat{p}_j,\widehat{\gamma}_j}(y)$,

$$\begin{aligned}
\ln h_{\beta_j}(x,y) \;=\; & \mathbb{I}(x=0, y=0)\, \ln C^{\chi}_{\frac{\beta_j}{1+\beta_j^2},\widehat{a}}(\widehat{p}_j, \widehat{p}_j) \\
& + \mathbb{I}(x=0, y>0)\, \ln C^{\chi,01}_{\frac{\beta_j}{1+\beta_j^2},\widehat{a}}(\widehat{p}_j, v) \\
& + \mathbb{I}(x>0, y=0)\, \ln C^{\chi,10}_{\frac{\beta_j}{1+\beta_j^2},\widehat{a}}(u, \widehat{p}_j) \\
& + \mathbb{I}(x>0, y>0)\, \ln c^{\chi}_{\frac{\beta_j}{1+\beta_j^2},\widehat{a}}(u, v).
\end{aligned}$$

Chapitre 4

Illustrations sur des données autour de la Baie de San Francisco

4.1 Présentation des données et analyses préliminaires

Ce chapitre permettra d'illustrer, sur de vraies données, l'utilité des nouveaux modèles spatio-temporels introduits au Chapitre 3. Les données qui seront étudiées ont été recueillies par le Service météorologique des États-Unis (le *National Weather Service*), une branche de l'agence gouvernementale *National Oceanic and Atmospheric Administration*. Cette dernière est chargée de la recherche et de la surveillance de l'atmosphère et des océans sur l'ensemble du territoire américain.

Plus précisément, les données considérées concernent douze stations d'observations situées autour de la Baie de San Francisco, en Californie. Cette portion du territoire possède un climat unique aux États-Unis, notamment par la présence de plus ou moins longues périodes de sécheresse. L'emplacement de chacune des douzes stations est identifié sur la carte de la Figure 4.1.

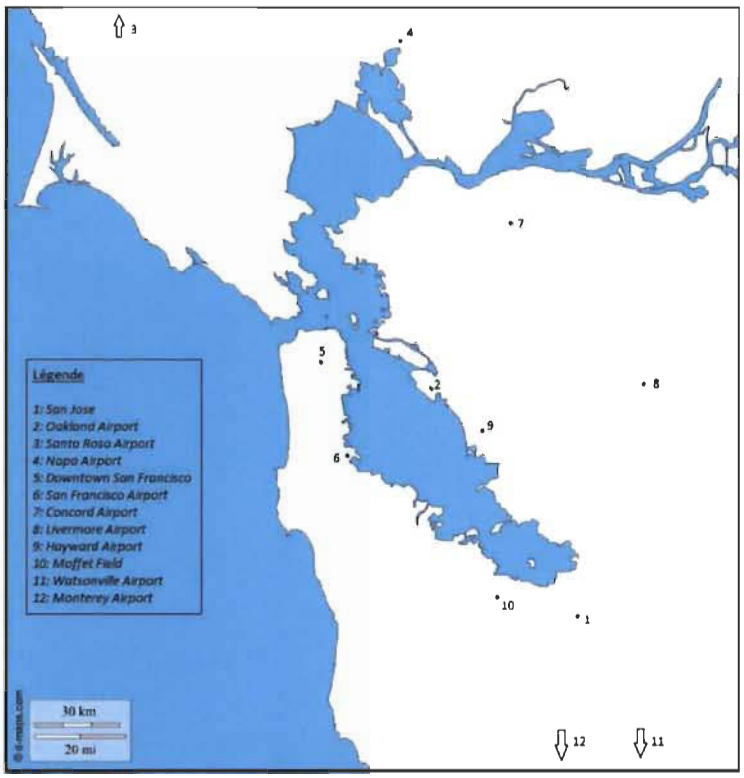

FIGURE 4.1 – Emplacement des douze stations d'observations autour de la Baie de San Francisco

L'aspect fondamental des modèles développés dans ce mémoire concerne le fait que les niveaux de dépendance entre les sites est fonction de la distance qui les sépare. Dans le cas des stations de la Figure 4.1, les distances se calculent à partir des coordonnées de longitude et de latitude. En effet, en tenant compte que la Terre présente une forme sphérique, on peut calculer l'orthodromie, c'est-à-dire le chemin le plus court entre deux points d'une sphère (distance à vol d'oiseau), à partir des coordonnées sphériques exprimées en termes de latitude et de longitude. Pour être plus précis, soient deux points A et B sur une sphère dont les coordonnées sont respectivement $(A^{\text{Lat}}, A^{\text{Long}})$ et $(B^{\text{Lat}}, B^{\text{Long}})$. Alors l'orthodromie entre A et B se calcule avec la formule

$$D(A,B) = R \cos^{-1}\left\{\sin\left(A^{\text{Lat}}\right)\sin\left(B^{\text{Lat}}\right) + \cos\left(A^{\text{Lat}}\right)\cos\left(B^{\text{Lat}}\right)\cos\left(B^{\text{Long}} - A^{\text{Long}}\right)\right\},$$

où $R = 6\ 371$ km est le rayon terrestre.

Dans la suite, trois séries chronologiques à $d = 12$ dimensions, notées $\mathcal{J}_1$, $\mathcal{J}_2$ et $\mathcal{J}_3$, seront considérées pour les analyses :

($\mathcal{J}_1$) Les maxima quotidiens de températures (en degrés Fahrenheit) aux mois de janvier, février, mars et avril 2017, pour un total de $n = 120$ observations ;

($\mathcal{J}_2$) Les températures moyennes mensuelles (en degrés Fahrenheit) de mai 2012 à avril 2017, pour un total de $n = 60$ observations ;

($\mathcal{J}_3$) Les totaux mensuels de précipitations (en pouces) entre avril 2012 et mars 2017, pour un total de $n = 60$ observations.

Comme on le verra, les observations du jeu de données $\mathcal{J}_1$ sont non stationnaires ; on travaillera alors avec la série différentiée d'ordre un, c'est-à-dire $\widetilde{X}_{tj} = X_{tj} - X_{t-1,j}$. Pour ce qui est de $\mathcal{J}_2$, on considérera plutôt la série *désaisonnalisée* $\widetilde{X}_{tj} = X_{tj} - X_{t-12,j}$ afin d'éliminer un effet significatif des saisons sur les observations. Quelques statistiques descriptives pour ces données se retrouvent d'ailleurs dans le Tableau 4.1. À noter enfin que l'estimation des paramètres sera basée sur une fraction de l'ensemble des 66 paires de sites, à l'image de l'estimateur de l'Équation (3.7). Plus précisément, on considérera les trois plus proches voisins pour chacun des $d = 12$ sites.

TABLE 4.1 – Moyenne ($\bar{X}_n$) et écart-type (S_n) pour les douze stations des jeux de données $\mathcal{J}_1$, $\mathcal{J}_2$ et $\mathcal{J}_3$; pour $\mathcal{J}_3$: proportions de jours sans précipitations ($\widehat{p}_n$)

	Données $\mathcal{J}_1$		Données $\mathcal{J}_2$		Données $\mathcal{J}_3$		
Station	$\bar{X}_n$	S_n	$\bar{X}_n$	S_n	$\bar{X}_n$	S_n	$\widehat{p}_n$
1. San Jose	0,193	4,313	0,263	2,961	1,404	1,734	0,250
2. Oakland	0,177	3,924	0,456	3,071	1,920	2,310	0,267
3. Santa Rosa	0,227	5,386	0,048	2,801	3,186	4,602	0,167
4. Napa	0,210	4,268	0,298	2,875	1,986	2,812	0,233
5. SF Downtown	0,168	4,749	0,246	2,973	1,951	2,755	0,117
6. SF Airport	0,168	4,249	0,381	3,285	2,142	2,657	0,317
7. Concord	0,252	4,842	0,510	3,121	1,722	2,107	0,267
8. Livermore	0,261	4,877	0,015	3,447	1,681	2,065	0,283
9. Hayward	0,202	3,903	0,425	3,358	1,376	1,711	0,267
10. Moffet Field	0,193	4,267	0,185	3,020	1,387	1,770	0,250
11. Watsonville	0,177	4,515	0,200	3,743	2,042	3,081	0,167
12. Monterey	0,143	4,958	0,258	3,206	1,499	2,155	0,083

4.2 Données $\mathcal{J}_1$: maxima quotidiens de température

4.2.1 Modélisation des marges

Le principal atout de l'utilisation des copules pour la modélisation multidimensionnelle est la possibilité de choisir des marges appropriées, indépendamment de la structure de dépendance. La première étape de modélisation consiste donc à ajuster des modèles pour les lois marginales. Avant de procéder, il faut toutefois s'assurer que le processus observé soit stationnaire. En examinant les séries présentées à la Figure 4.2, on voit clairement qu'il y a une tendance à la hausse en fonction du temps.

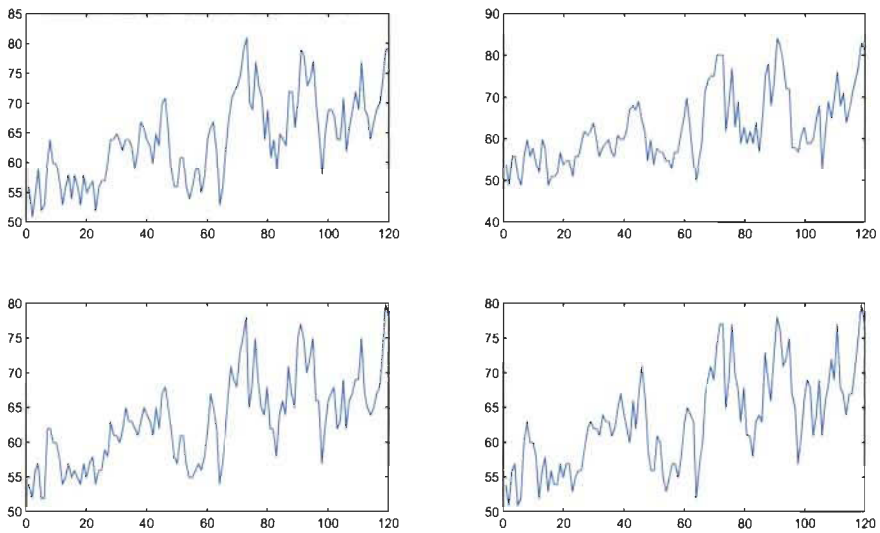

FIGURE 4.2 – De gauche à droite et de bas en haut : séries chronologiques des maxima quotidiens de températures aux stations San Jose, Santa Rosa, Hayward et Moffet Field

Pour rendre les séries stationnaires, on considérera les transformations différentielles d'ordre un, c'est-à-dire $\widetilde{X}_{tj} = X_{tj} - X_{t-1,j}$. En regardant les séries transformées à la Figure 4.3, on constate que l'opération a fonctionné en ce sens que les données résultantes sont stationnaires ; dans la suite, on travaillera donc avec ces séries.

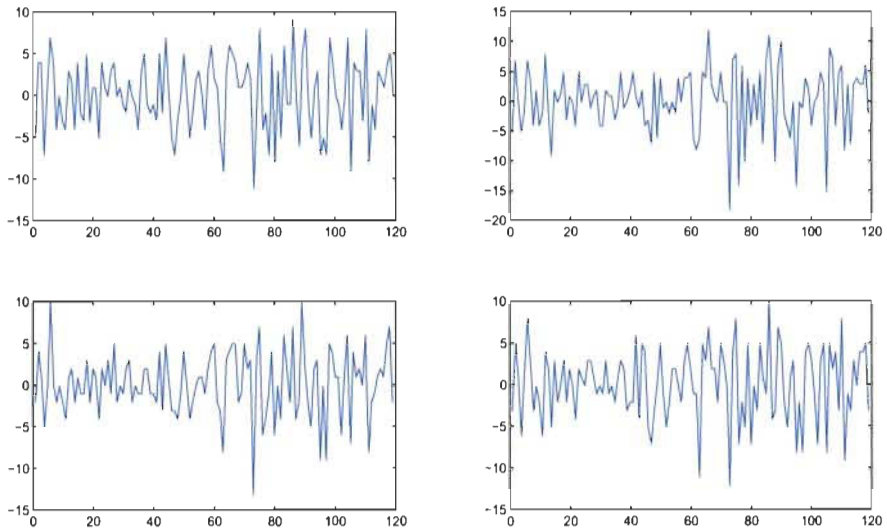

FIGURE 4.3 – De gauche à droite et de bas en haut : séries chronologiques des différences d'ordre un des maxima quotidiens de tempéatures aux stations San Jose, Santa Rosa, Hayward et Moffet Field

Les différences des maxima de températures sont distribuées selon une loi continue. On peut voir de quoi ont l'air ces distributions en regardant la diagonale de la Figure 4.4, qui contient les histogrammes des stations San Jose, Santa Rosa, Hayward et Moffet Field. À la lumière de ces histogrammes, qui sont représentatifs de ceux des huit autres stations, on pourrait considérer des lois Normales pour toutes les marges. Ainsi, les estimations des paramètres se déduisent des informations du Tableau 4.1 sur les moyennes et les écart-types.

4.2.2 Modélisation de la dépendance spatiale

Maintenant, conditionnellement aux estimations des marges, il reste à caractériser la dépendance spatiale entre les stations. Sous l'hypothèse d'une copule qui appartient à la famille des modèles Khi-deux, cela revient à estimer les paramètres θ, ϵ et a. La Figure 4.4 présente les nuages de points pour les stations San Jose, Santa Rosa, Hayward et Moffet Field. Pour bien les interpréter, il faut savoir que les stations San

Jose et Moffet Field sont près l'une de l'autre ; en revanche, elles sont très éloignées de la station Santa Rosa. Quand à elle, la station Hayward est relativement près des stations San Jose et Moffet Field, et assez éloignée de la station Santa Rosa.

La construction des nuages de points du triangle inférieur de la Figure 4.4 est basés sur les rangs des observations divisés par n ; cela permet de visualiser la forme des copules pour les paires considérées. À première vue, les copules de ces paires semblent en symétrie radiale. En effet, les queues inférieures et supérieures semblent très similaires. On s'attend donc à une estimation élevée du paramètre a de décentralité, ce qui correspondrait, à toute fin pratique, au choix de la copule Normale.

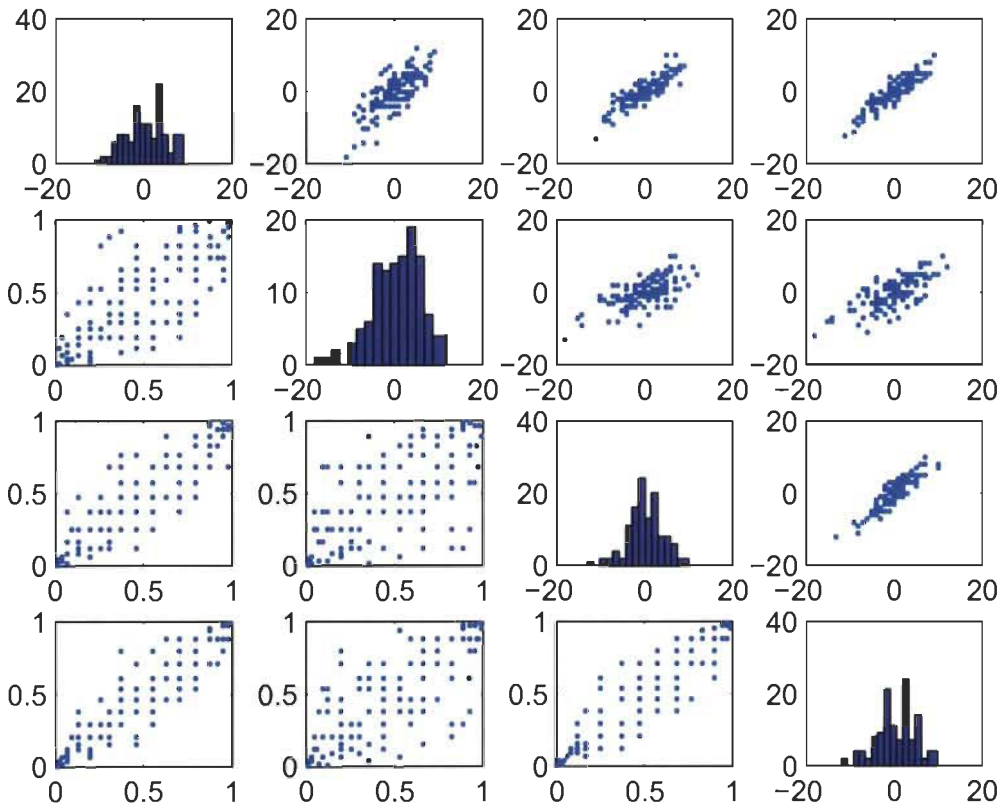

FIGURE 4.4 – Dépendance spatiale entre les stations San Jose, Santa Rosa, Hayward et Moffet Field pour les maxima quotidiens de températures ; diagonale : histogrammes ; triangle supérieure : nuages de points ; triangle inférieure : copule empirique

Une hypothèse fondamentale quant à la modélisation spatiale avec les outils du Cha-

pitre 3 est l'existence d'une relation entre les niveaux de dépendance des paires et la distance qui les sépare ; on parle de l'hypothèse d'isotropie. L'évolution des tau de Kendall en fonction des distances est présenté sur le graphique de gauche de la Figure 4.5. On observe une nette décroissance des tau de Kendall en fonction de la distance, telle que requise par l'hypothèse d'isotropie.

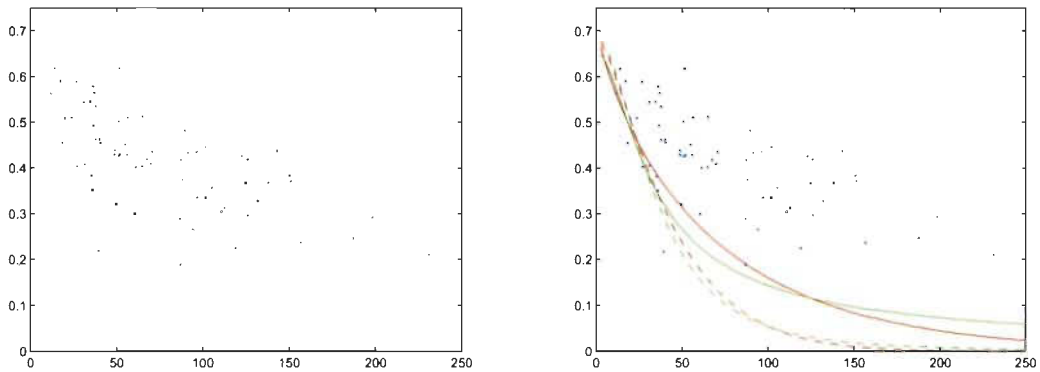

FIGURE 4.5 – À gauche : tau de Kendall en fonction de la distance (en km) pour toutes les paires du jeu de données sur les maxima quotidiens de températures ; à droite : courbes du tau de Kendall pour la fonction Matérn (rouge) et Rationnelle Quadratique (vert) lorsque $\nu = 1/2$ (trait continu) et $\nu = 3/2$ (trait discontinu)

Les résultats des estimations des paramètres θ, ϵ et a se trouvent au Tableau 4.2. En se basant sur ces estimations, on peut tracer la courbe de l'évolution du tau de Kendall en fonction de la distance pour chacun des quatre modèles considérés. On peut voir le résultat sur le graphique de droite de la Figure 4.5. On constate que les quatre modèles ont de la difficulté à bien reproduire le comportement du tau de Kendall pour des distances relativement grandes. Néanmoins, pour des distances inférieures à 50 km, les ajustements sont relativements bons. Globalement, un choix adéquat serait la fonction Matérn de paramètre $\nu = 1/2$.

TABLE 4.2 – Résultats des estimations des paramètres de dépendance spatiale θ, ϵ et a pour les jeux de données $\mathcal{J}_1$, $\mathcal{J}_2$ et $\mathcal{J}_3$

Fonction de lien	Données $\mathcal{J}_1$			Données $\mathcal{J}_2$			Données $\mathcal{J}_3$		
	$\widehat{\theta}$	$\widehat{\epsilon}$	$\widehat{a}$	$\widehat{\theta}$	$\widehat{\epsilon}$	$\widehat{a}$	$\widehat{\theta}$	$\widehat{\epsilon}$	$\widehat{a}$
$g_{1/2}^{\text{Mat}}$	78,42	0,100	2,11	666,72	0,059	4,55	160.34	0.001	1.64
$g_{3/2}^{\text{Mat}}$	25,90	0,113	1,68	151,97	0,047	0,33	143.13	0.018	1.33
$g_{1/2}^{\text{RQ}}$	27,38	0,138	1,94	225,65	0,046	0,47	149.80	0.016	1.29
$g_{3/2}^{\text{RQ}}$	52,96	0,106	1,62	1 371,90	0,075	2,26	170.09	0.006	1.41

4.2.3 Reproduction du phénomène selon le modèle choisi

Un des objectifs de l'ajustement d'un modèle à un phénomène naturel est la possibilité de le reproduire. C'est ce que nous allons faire ici en tenant compte du modèle choisi. Tout juste avant, il est important de noter que pour ces données, il n'y a pas de dépendance sérielle significative, ce qui veut dire que l'étape d'estimation des paramètres de dépendance sérielle β n'est pas requise. Tel que mentionné précédemment, un modèle adéquat survient lorsque la fonction de lien est Matérn de paramètre $\nu = 1/2$, dans lequel cas on a $\widehat{\theta} = 78,42$, $\widehat{\epsilon} = 0,100$ et $\widehat{a} = 2,11$. Cette valeur élevée du paramètre de décentralité a indique une dépendance proche de la copule Normale, tel que pressenti lors de l'examen des nuages de points du triangle inférieur de la Figure 4.4. À noter que la valeur du paramètre de portée qui correspond à ce modèle, tel que défini à la fin de la Section 3.2.5, est $\widehat{\mathcal{D}} = 83,4$ km.

La Figure 4.6 présente des données simulées aux stations San Jose, Santa Rosa, Hayward et Moffet Field en considérant les paramètres estimés des marges et de la dépendance spatiale. On peut constater des similitudes avec la Figure 4.4, qui est basée sur les vraies données. On note toutefois que globalement, les niveaux de dépendance

sont légèrement sous-estimés, ce qui est cohérent avec le graphique de droite de la Figure 4.5. Néanmoins, les principales caractéristiques du phénomène telles les lois marginales et la symétrie radiale, sont adéquatement reproduites par le modèle.

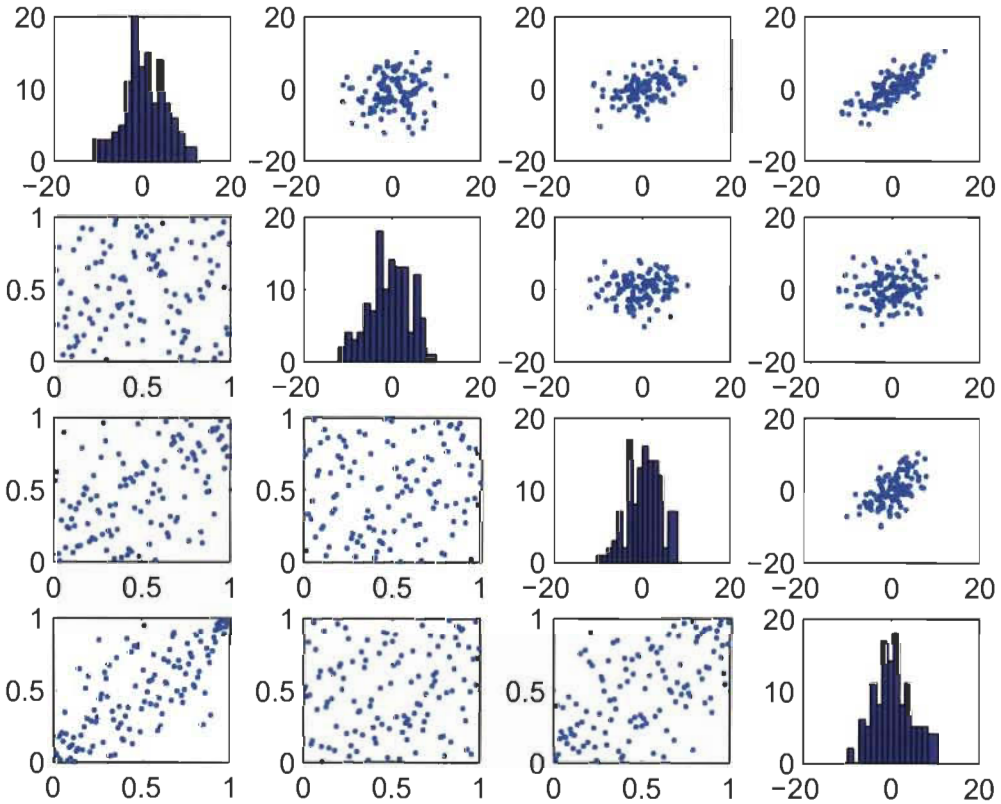

FIGURE 4.6 – Reproduction de la dépendance spatiale entre les stations San Jose, Santa Rosa, Hayward et Moffet Field basée sur des données simulées

4.3 Données $\mathcal{J}_2$: températures moyennes mensuelles

4.3.1 Modélisation des marges

Tout comme les données sur les maxima de températures, les températures moyennes mensuelles sont sujettes à une forte absence de stationnarité. En effet, comme on peut

le constater à la Figure 4.7, on voit qu'il y a une forte saisonnalité d'ordre douze. Pour stabiliser les séries, on va donc travailler plutôt avec les différences $\widetilde{X}_{tj} = X_{tj} - X_{t-12,j}$; la Figure 4.8 montre que cette opération a réussi à rendre ces séries stationnaires. Le jeu de données comporte désormais $n = 48$ observations, cependant.

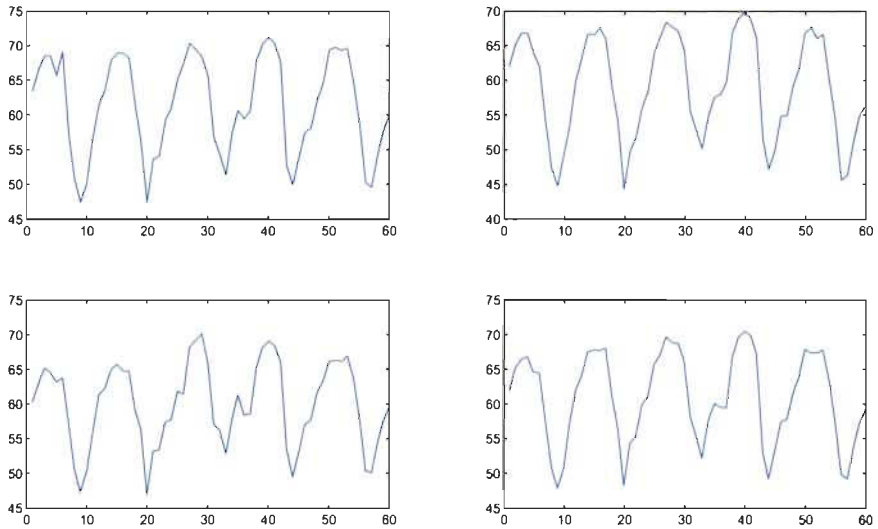

FIGURE 4.7 – Séries chronologiques des températures moyennes mensuelles aux stations San Jose, Santa Rosa, Hayward et Moffet Field

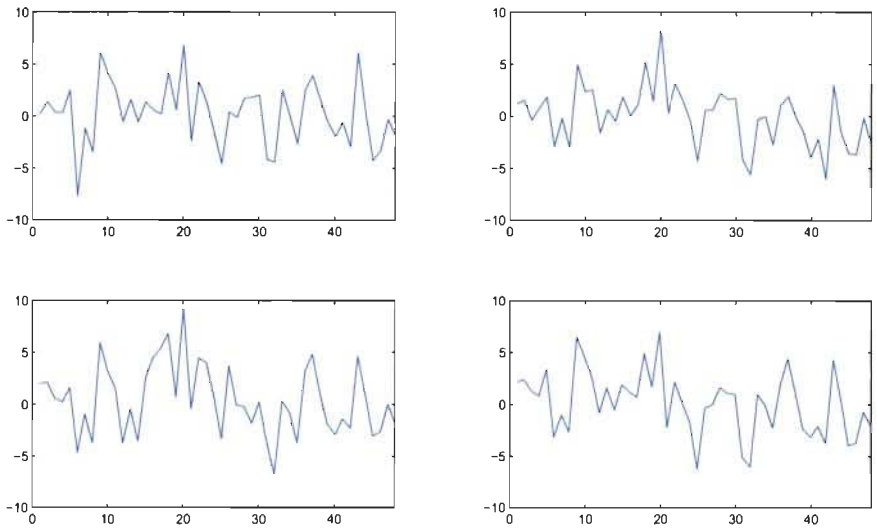

FIGURE 4.8 – Séries chronologiques des différences d'ordre douze des températures moyennes mensuelles aux stations San Jose, Santa Rosa, Hayward et Moffet Field

Les histogrammes pour les stations San Jose, Santa Rosa, Hayward et Moffet Field se

retrouvent sur la diagonale de la Figure 4.9 ; ceux-ci sont représentatifs des huit autres stations d'observations. En regardant ces histogrammes, il semble raisonnable de supposer que les lois marginales appartiennent à la famille des distributions Normales $\mathbb{N}(\mu, \sigma^2)$. Dans ce cas, pour chaque station $j \in \{1, \ldots, 12\}$, les estimateurs à maximum de vraisemblance de μ_j et de σ_j^2 seront respectivement la moyenne et la variance empiriques pour cette station ; ces informations peuvent être tirées du Tableau 4.1.

4.3.2 Modélisation de la dépendance spatiale

Pour se donner une idée du type de dépendance qui existe entre les différentes stations, la Figure 4.9 présente les nuages de points pour les stations San Jose, Santa Rosa, Hayward et Moffet Field. On rappelle que les stations San Jose et Moffet Field sont rapprochées, mais éloignées de la station Santa Rosa, alors que Hayward est proche des stations San Jose et Moffet Field et assez éloignée de la station Santa Rosa. En examinant les nuages de points du triangle inférieur, on peut observer la forme de la dépendance entre les stations, c'est-à-dire la copule. De façon générale, les liens les plus forts s'observent pour des stations qui sont rapprochées. Aussi, pour chaque paire de stations, la dépendance semble symétrique selon la diagonale, tel que requis implicitement par le modèle. De plus, il semble y avoir une concentration relativement plus élevée de points dans les queues supérieures que dans les queues inférieures. Cette caractéristique devrait être bien captée par une copule Khi-deux.

L'évolution des tau de Kendall en fonction des distances est présenté au graphique de gauche de la Figure 4.10. À première vue, la tendance à la baisse des tau de Kendall souhaitée par le modèle semble plus ou moins valide. Un examen plus approfondi montre toutefois que si on fait abstraction de quelques points aberrants entre les distances 100 km et 150 km, l'hypothèse peut être soutenue. Il faut garder à l'esprit que le nombre d'observations est relativement faible ici, ce qui entraîne des estimations des tau de Kendall sujettes à d'assez grandes variabilités. Le Tableau 4.2 présente les

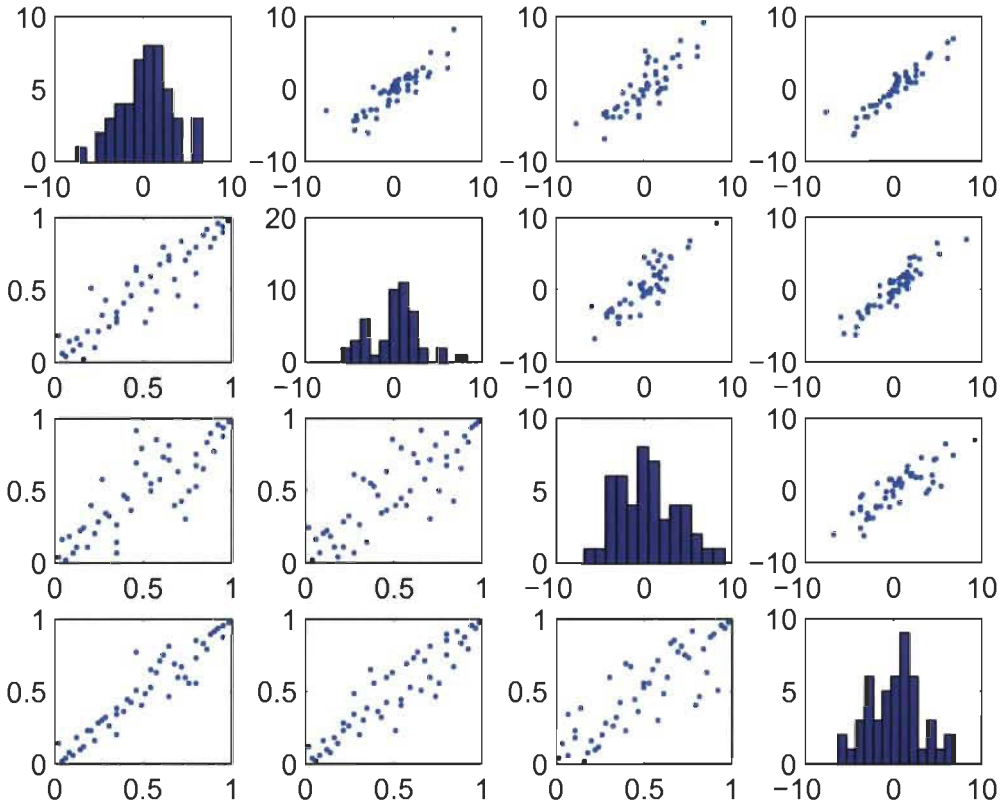

FIGURE 4.9 – Dépendance spatiale entre les stations San Jose, Santa Rosa, Hayward et Moffet Field pour les températures moyennes mensuelles ; diagonale : histogrammes ; triangle supérieure : nuages de points ; triangle inférieure : copule empirique

estimations des paramètres θ, ϵ et a pour les fonctions de lien Matérn et Rationnelle Quadratique lorque $\nu \in \{1/2, 3/2\}$. On voit qu'il y a énormément de disparité entre les valeurs estimées selon la fonction de lien considérée. Si on se fie au graphique de droite de la Figure 4.10, un choix approprié serait Matérn avec $\nu = 1/2$.

Avec ce choix de modèle, on pourrait reproduire le phénomène des températures moyennes mensuelles par simulations. Toutefois, contrairement aux maxima de température, il y a de la dépendance sérielle dans quelques-unes des séries individuelles. En fait, en examinant le graphique des valeurs des tau de Kendall de délai $k \in \{1, \ldots, 4\}$, on remarque une dépendance sérielle qui s'apparente à un processus de dépendance d'ordre un pour les stations Concord, Livermore, Moffet Field et Watsonwille. Les

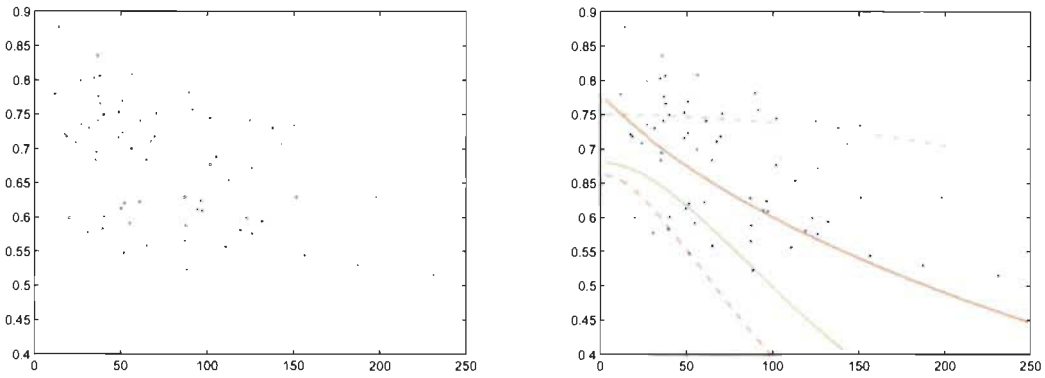

FIGURE 4.10 – À gauche : tau de Kendall en fonction de la distance (en km) pour toutes les paires du jeu de données sur les températures moyennes mensuelles ; à droite : courbes du tau de Kendall pour la fonction Matérn (rouge) et Rationnelle Quadratique (vert) lorsque $\nu = 1/2$ (trait continu) et $\nu = 3/2$ (trait discontinu)

estimations du paramètre sériel pour ces stations sont d'ailleurs respectivement $\widehat{\beta}_7 = 0,161$, $\widehat{\beta}_8 = 0,229$, $\widehat{\beta}_{10} = 0,038$ et $\widehat{\beta}_{11} = 0,048$; les valeurs estimées du paramètre sont très près de zéro pour les huit autres stations.

4.4 Données $\mathcal{J}_3$: totaux mensuels de précipitations

4.4.1 Modélisation des marges

La particularité des données de précipitations est qu'elles comportent habituellement un nombre important de zéros ; il faut en tenir compte dans les analyses. En se rapportant au Tableau 4.1, on voit que le jeu de données $\mathcal{J}_3$ contient effectivement des proportions significatives de zéros, ce qui appelle à utiliser la méthodologie des Sections 3.5 et 3.6 du mémoire. Voir également la Figure 4.11 concernant les séries chronologiques des stations San Jose, SF Downtown, SF Airport et Moffet field.

Comme première étape de modélisation, on va donc supposer des lois de la forme

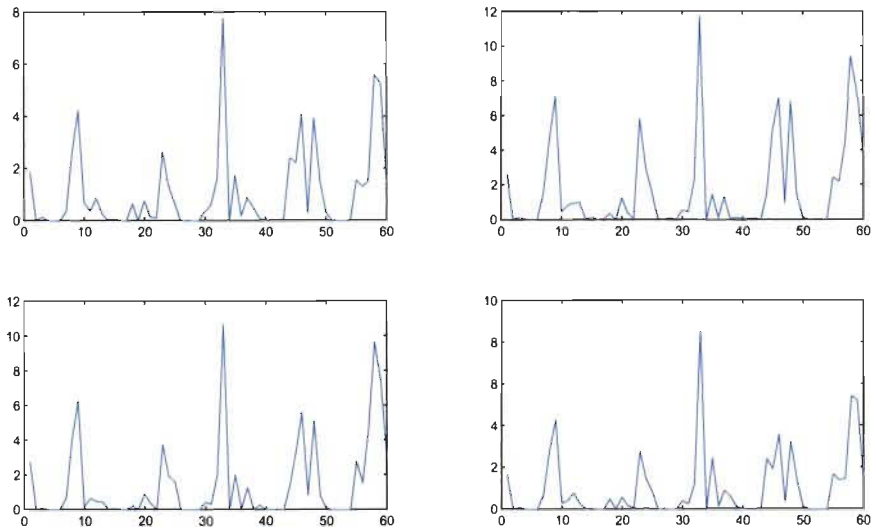

FIGURE 4.11 – De gauche à droite et de bas en haut : séries chronologiques des totaux de précipitations mensuelles aux stations San Jose, SF Downtown, SF Airport et Moffet field

$F_p(x) = p+(1-p)F(x)$ pour les fonctions de répartition marginales. Les histogrammes sur la diagonale de la Figure 4.12 permettent de penser que des lois Exponentielles sont appropriées pour la modélisation de la partie continue de la distribution. Une autre confirmation est donnée par les valeurs des moyennes et écart-types que l'on retrouvent au Tableau 4.1. En effet, celles-ci sont relativement identiques, ce qui est le cas théoriquement pour la distribution Exponentielle. L'estimation des paramètres $\gamma_1, \ldots, \gamma_{12}$ sera donc basée sur les moyennes empiriques présentées au Tableau 4.1. Les valeurs estimées de $p_1, \ldots, p_d$ sont simplement les proportions observées de zéros.

4.4.2 Modélisation de la dépendance spatiale

La Figure 4.12 montre les nuages de points pour les stations San Jose, SF Downtown, SF Airport et Moffet field. En faisant abstraction des zéros, on peut penser que la dépendance se modélise adéquatement avec une copule Khi-deux. Il semble y avoir, en outre, une certaine asymétrie radiale qui devrait se réfléter dans la valeur estimée

du paramètre de décentralité. Le graphique de gauche de la Figure 4.13 montre le comportement du tau de Kendall en fonction des distances. L'hypothèse d'isotropie tient la route, même si quelques points semblent s'éloigner de la tendance générale.

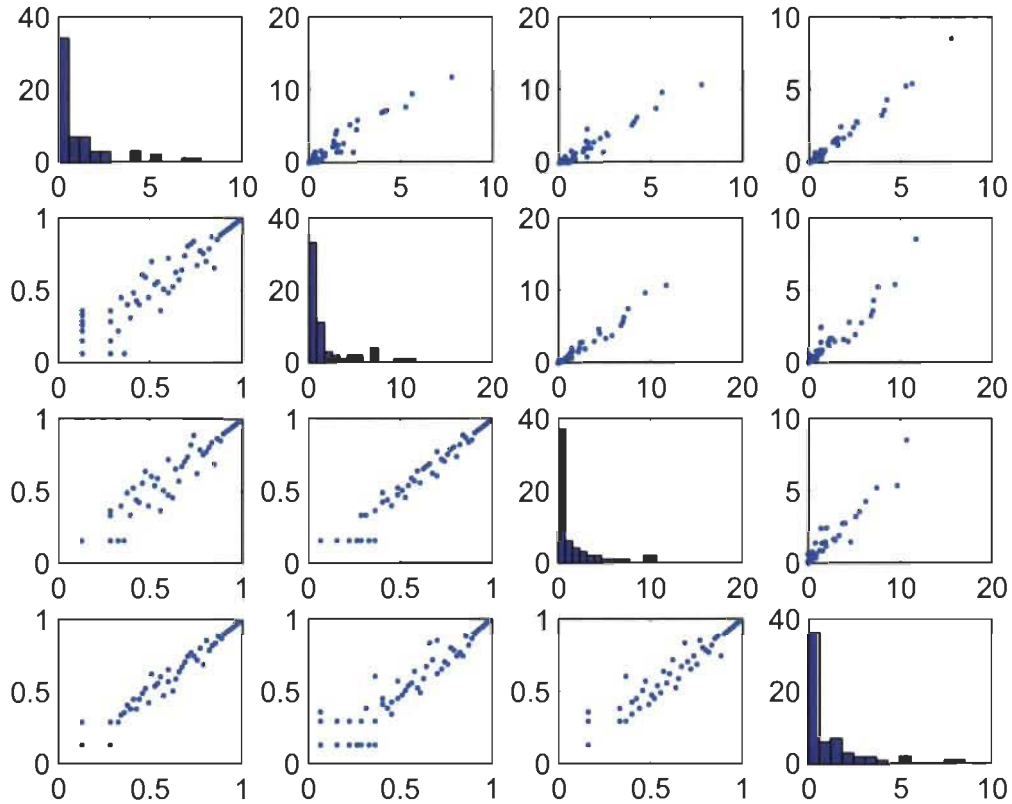

FIGURE 4.12 – Dépendance spatiale entre les stations San Jose, SF Downtown, SF Airport et Moffet field pour les totaux de précipitations mensuelles ; diagonale : histogrammes ; triangle supérieure : nuages de points ; triangle inférieure : copule empirique

Le Tableau 4.2 présente les estimations des paramètres θ, ϵ et a pour les fonctions de lien Matérn et Rationnelle Quadratique lorque $\nu \in \{1/2, 3/2\}$. Pour ce faire, on a utilisé la fonction de vraisemblance de l'Équation (3.12). Ces estimations sont assez semblables pour les quatre fonctions de lien. On voit que l'effet de pépite est à toutes fins utiles nul, alors que le paramètre a de décentralité est légèrement supérieur à 1 ; le modèle parvient donc à capturer l'asymétrie radiale présente dans ces données. À la lumière du graphique de droite de la Figure 4.13, les fonctions Matérn avec $\nu = 3/2$ et Rationnelle Quadratique avec $\nu = 1/2$ réussissent à bien reproduire la dépendance

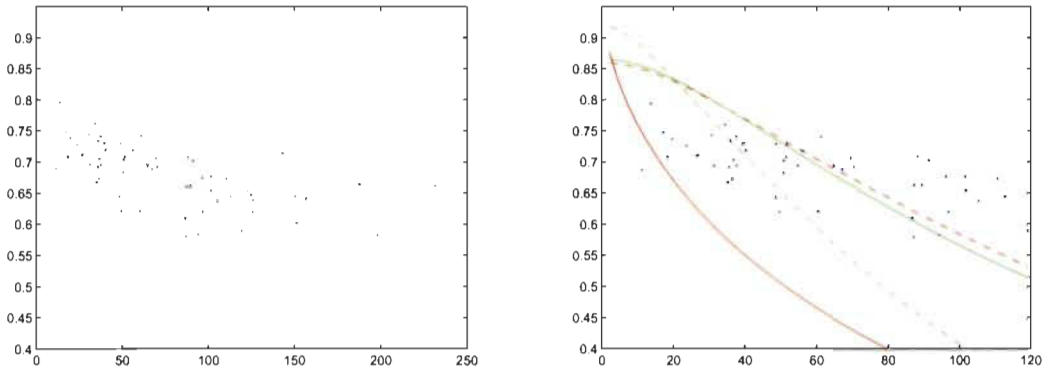

FIGURE 4.13 – À gauche : tau de Kendall en fonction de la distance (en km) pour toutes les paires du jeu de données sur les totaux de précipitations mensuelles ; à droite : courbes du tau de Kendall pour la fonction Matérn (rouge) et Rationnelle Quadratique (vert) lorsque $\nu = 1/2$ (trait continu) et $\nu = 3/2$ (trait discontinu)

en fonction de la distance entre les stations.

4.4.3 Reproduction du phénomène selon le modèle choisi

Tel que mentionné, un modèle adéquat pour la dépendance spatiale est obtenu lorsque la fonction de lien est Rationnelle Quadratique de paramètre $\nu = 1/2$, dans lequel cas on a $\widehat{\theta} \approx 150$, $\widehat{\epsilon} \approx 0$ et $\widehat{a} \approx 1,3$. Le paramètre de portée sous ce modèle est estimé par $\widehat{\mathcal{D}} = 494,77$ km, ce qui montre qu'il y a beaucoup de dépendance dans ce jeu de données. En effet, comme la distance entre deux stations est au maximum 250 km, toutes les paires de sites ont un tau de Kendall supérieur à 0,2 selon le modèle choisi.

Afin de simuler le phénomène des précipitations autour de la Baie de San Francisco, il reste à évaluer la dépendance sérielle sous l'hypothèse d'un modèle sous-jacent à moyenne mobile d'ordre un. Les estimations des paramètres pour chacune des stations sont $\widehat{\beta}_1 = 0,476$, $\widehat{\beta}_2 = 0,520$, $\widehat{\beta}_3 = 0,480$, $\widehat{\beta}_4 = 0,452$, $\widehat{\beta}_5 = 0,504$, $\widehat{\beta}_6 = 0,615$, $\widehat{\beta}_7 = 0,468$, $\widehat{\beta}_8 = 0.615$, $\widehat{\beta}_9 = 0,422$, $\widehat{\beta}_{10} = 0,435$, $\widehat{\beta}_{11} = 0.631$ et $\widehat{\beta}_{12} = 0,535$.

La Figure 4.14 présente des données simulées aux stations San Jose, SF Downtown, SF Airport et Moffet field en considérant les paramètres estimés des marges, de la dépendance spatiale et de la dépendance sérielle. On peut constater la ressemblance frappante entre le comportement de ces données simulées et ce celui des vraies données tel qu'observé à la Figure 4.12. On a donc réussi à reproduire le phénomène des précipations de façon satisfaisante.

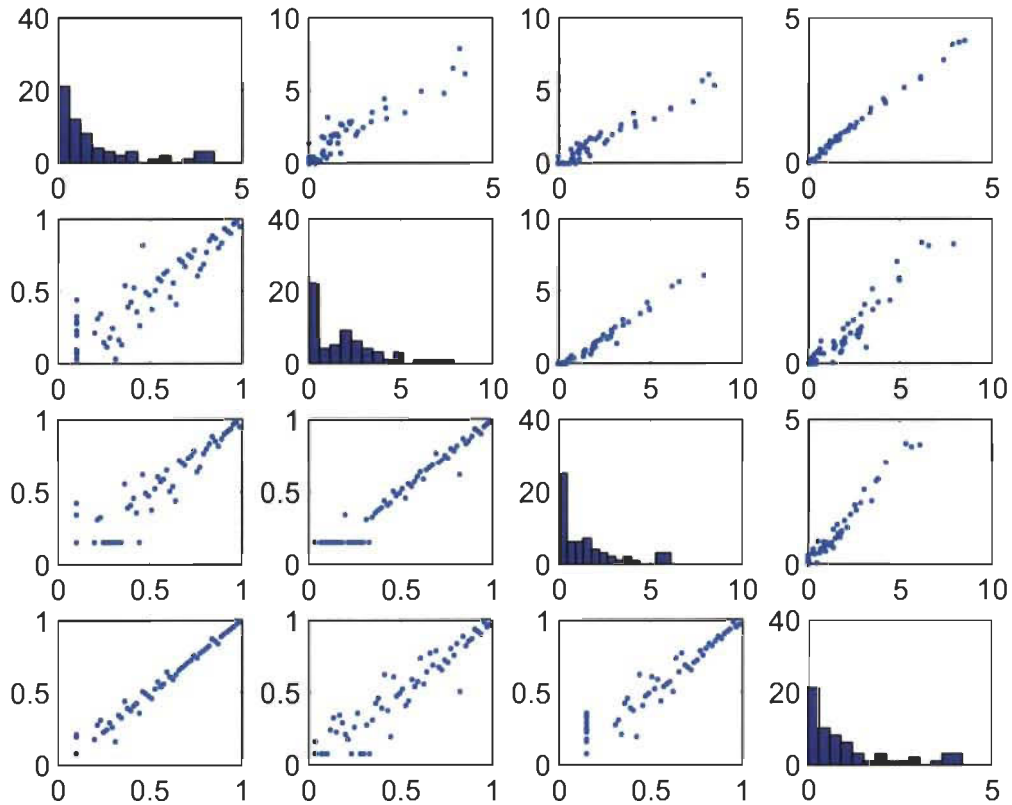

FIGURE 4.14 – Reproduction de la dépendance spatiale entre les stations San Jose, SF Downtown, SF Airport et Moffet field basée sur des données simulées

Conclusion

Dans ce mémoire, de nouveaux modèles spatio-temporels basés sur les champs aléatoires Khi-deux ont été introduits et étudiés. Dans un premier temps, la famille des copules Khi-deux a été présentée, autant pour le cas bivarié que multidimensionnel. L'étude rigoureuse de cette copule a mené à plusieurs propriétés utiles lors de l'utilisation de celle-ci, notamment la possibilité de contenir de l'asymétrie radiale. C'est ce qui en fait, notamment, une alternative très intéressante à la copule Normale, qui est beaucoup utilisée. L'estimation de ses paramètres a aussi été explorée.

La branche particulière de la modélisation spatiale a été introduite et par la suite, adaptée pour inclure la théorie des copules, ce qui permet de lier ensemble les deux principaux domaines étudiés dans ce mémoire. Ceci a mené à la construction de quatre différents modèles basés sur la copule Khi-deux, chacun possédant des caractéristiques distinctes quant à l'inclusion de composantes temporelle, sérielle et discrète. Des estimateurs propres à ces modèles pour chacun de ses paramètres ont été suggérés et validés par une étude de performance par simulations et ce, pour plusieurs scénarios possibles incluant les copules Khi-deux et Normale couplées à différentes fonctions de lien et des valeurs variantes des paramètres.

Pour terminer, les nouveaux modèles spatio-temporels proposés ont été mis en application pour modéliser des précipitations et des températures dans les alentours de la baie de San Francisco. Une étude préliminaire sur différents aspects des jeux de

données, notamment sur les distributions marginales et la dépendance spatiale, nous a permis de procéder à un ajustement maximal des modèles pour les données de ces phénomènes météorologiques possédant des caractéristiques propres.

Plusieurs éléments omis dans ce mémoire mériteraient d'être étudiés dans d'éventuels projets de recherches. En effet, le domaine étudié est si vaste et rempli de possibilités qu'il pourrait facilement faire l'objet de recherches subséquentes. Notons premièrement l'étude de la performance des estimateurs des nouveaux modèles par simulations. Il serait intéressant de considérer des cas différents de la copule Khi-deux autres que le cas centré pour maximiser son efficacité et découvrir des situations où son utilisation est nettement plus avantageuse que celle de la copule Normale. D'autre part, il pourrait être intéressant d'étendre nos horizons pour ce qui est du modèle spatio-temporel pour le cas sériel avec marges continues. Nous l'avons étudié pour un type de dépendance sérielle très précise avec le modèle à moyenne mobile d'ordre un. Son adaptation pour d'autres types spécifiques de dépendance sérielle serait une expansion démontrant sa grande variété. Il serait aussi pertinent d'étudier le processus d'interpolation spatiale avec les nouveaux modèles spatio-temporels introduits. Il s'agirait d'un défi plus qu'intéressant qui viendrait bonifier la raison d'être de ces derniers.

Bibliographie

[1] Andràs Bàrdossy. Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, 42(11) :1–12, 2006.

[2] András Bárdossy and Jing Li. Geostatistical interpolation using copulas. *Water Resources Research*, 44(7) :W07412, 2008.

[3] D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1) :141–151, 1978.

[4] Noel Cressie. *Statistics for spatial data*. John Wiley and Sons, Inc, 1993.

[5] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3) :543–552, 1995.

[6] B. Gräler and E. Pebesma. The pair-copula construction for spatial data : a new approach to model spatial dependency. *Procedia Environmental Sciences*, 7 :206–211, 2011.

[7] Nils Lid Hjort. A quasi-likelihood method for estimating parameters in spatial covariance functions. Technical report, Technical Report SAND/93, Norwegian Computing Centre, Oslo, 1993.

[8] Harry Joe. *Dependence modeling with copulas*, volume 134 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[9] Cari Kaufman and Benjamin Shaby. The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2) :473–484, 2013.

[10] H. Kazianka and J. Pilz. Bayesian spatial modeling and interpolation using copulas. *Computers & Geosciences*, 37(3) :310–319, 2011.

[11] Hannes Kazianka and Jürgen Pilz. Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*, 24(5) :661–673, 2010.

[12] D. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society*, 52 :119–139, 1951.

[13] Jing Li. *Application of copulas as a new geostatistical tool.* PhD thesis, Stuttgart University, Allemagne, 2010.

[14] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management : Concepts, techniques and tools.* Princeton University Press, Princeton, NJ, 2005.

[15] Christian Meyer. The bivariate normal copula. *Comm. Statist. Theory Methods*, 42(13) :2402–2422, 2013.

[16] Roger B. Nelsen. *An introduction to copulas.* Springer Series in Statistics. Springer, New York, second edition, 2006.

[17] Roger B. Nelsen. Extremes of nonexchangeability. *Statist. Papers*, 48 :329–336, 2007.

[18] Jean-François Quessy, Louis-Paul Rivest, and Marie-Hélène Toupin. Semiparametric pairwise inference methods in spatial models based on copulas. *Spat. Stat.*, 14(part B) :472–490, 2015.

[19] Jean-François Quessy, Louis-Paul Rivest, and Marie-Hélène Toupin. On the family of multivariate chi-square copulas. *J. Multivariate Anal.*, 152 :40–60, 2016.

[20] Brian D. Ripley. *Spatial statistics.* John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[21] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8 :229–231, 1959.

[22] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statist. Sinica*, 21(1) :5–42, 2011.

[23] Cristiano Varin and Paolo Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3) :519–528, 2005.

[24] G. N. Watson. *A treatise on the theory of Bessel functions*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1995. Reprint of the second (1944) edition.

Annexe A

Démonstrations de résultats originaux

A.1 Preuve du Lemme 3.1

Supposons que $F_{p,\gamma}^{-1}(u) \leq x$. Alors puisque

$$F_{p,\gamma}^{-1}(u) = F_{\gamma}^{-1}\left\{\frac{\max(u-p,0)}{1-p}\right\},$$

on obtient $F_{p_\gamma}^{-1}(u) \leq x$ si et seulement si $\max(u,p) \leq F_{p,\gamma}$. Ceci permet de conclure que $u \leq \max(u,p) \leq F_{p,\gamma}(x)$. Maintenant, si on suppose que $u \leq F_{p,\gamma}(x)$, cela revient à écrire $(u-p)/(1-p) \leq F_{\gamma}(x)$. En appliquant F_{γ} de chaque côté de cette inégalité, on a du fait que F_{γ} est continue que

$$F_{\gamma}^{-1}\left(\frac{u-p}{1-p}\right) \leq x.$$

Par conséquent,

$$F_{p,\gamma}^{-1}(u) = F_{\gamma}^{-1}\left\{\frac{\max(u-p,0)}{1-p}\right\} \leq F_{\gamma}\left(\frac{u-p}{1-p}\right) \leq x.$$

A.2 Preuve de la Proposition 3.1

La fonction de log-vraisemblance pour $X_1, \ldots, X_n$ i.i.d. $F_{p,\gamma}$ est

$$\mathcal{L}(p,\gamma) = \sum_{t=1}^{n} \mathrm{I}(X_t = 0)\, \ln p + \sum_{t=1}^{n} \mathrm{I}(X_i > 0)\, \{\ln(1-p) + \ln f_\gamma(X_t)\}\,.$$

En dérivant par rapport à p, on obtient

$$\begin{aligned}
\frac{\partial}{\partial p}\mathcal{L}(p,\gamma) &= \frac{1}{p}\sum_{t=1}^{n} \mathrm{I}(X_t = 0) - \frac{1}{1-p}\sum_{t=1}^{n} \mathrm{I}(X_t > 0) \\
&= \frac{1}{p(1-p)}\left\{(1-p)\sum_{t=1}^{n} \mathrm{I}(X_t = 0) - p\sum_{t=1}^{n} \mathrm{I}(X_t > 0)\right\} \\
&= \frac{1}{p(1-p)}\left\{\sum_{t=1}^{n} \mathrm{I}(X_t = 0) - np\right\}.
\end{aligned}$$

La solution $\partial\mathcal{L}(p,\gamma)/\partial p = 0$ amène directement

$$\widehat{p} = \frac{1}{n}\sum_{t=1}^{n} \mathrm{I}(X_t = 0).$$

Maintenant, en dérivant rapport à γ, on trouve

$$\frac{\partial}{\partial \gamma}\mathcal{L}(p,\gamma) = \sum_{t=1}^{n} \mathrm{I}(X_t > 0)\, \frac{\partial}{\partial \gamma}\, \ln f_\gamma(X_t).$$

Autrement dit, $\widehat{\gamma}$ est l'estimateur à maximum de vraisemblance basé sur les $X_t > 0$.

Annexe B

Démonstrations de résultats de [19] concernant la copule Khi-deux

B.1 Preuve du Lemme 2.1

Premièrement, du fait que $G_{a_j}(x) = \Phi(\sqrt{x} - a_j) + \Phi(\sqrt{x} + a_j) - 1$, on trouve que $G_0(u) = 2\Phi(\sqrt{x}) - 1$, d'où $G_0^{-1}(u) = \left\{\Phi^{-1}\left((\frac{u+1}{2})\right)\right\}^2$. Alors, on obtient

$$h_0(u) = \text{sign}(u)\sqrt{G_0^{-1}(|u|)} = \Phi^{-1}\left(\frac{u+1}{2}\right) \quad \text{pour} \quad u \in [-1, 1].$$

Donc, sachant que $\tilde{h}_a(u) = \Phi(h_a(u))$, on a que $\tilde{h}_0(u) = \Phi(h_0(u)) = (1 + u)/2$ pour $u \in [-1, 1]$.

B.2 Preuve du Lemme 2.2

La démonstration découle directement de la représentation stochastique

$$\left\{G_{a_1}\left\{(Z_1+a_1)^2\right\}, G_{a_2}\left\{(Z_2+a_2)^2\right\}\right\} \sim C^{\chi}_{\rho,a_1,a_2},$$

où $(Z_1, Z_2) \sim \Phi_\rho$ et du fait que $(-Z_1, Z_2) \sim \Phi_{-\rho}$ et $(-Z_1, -Z_2) \sim \Phi_\rho$.

B.3 Preuve de la Proposition 2.1

Selon [15], la copule Normale bivariée peut s'écrire en incluant une intégrale sous la forme

$$C^{\mathrm{N}}_\rho(u_1, u_2) = u_1 u_2 + \int_0^\rho \phi_r\left\{\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right\} \mathrm{d}r.$$

En ajustant ce résultat, la copule Khi-deux centrée peut s'écrire comme étant

$$C^{\chi}_\rho(u_1, u_2) = u_1 u_2 + 2\int_0^\rho \{\phi_r(w_1, w_2) - \phi_r(-w_1, w_2)\}\mathrm{d}r,$$

où $w_1 = \Phi^{-1}\{(u_1+1)/2\} \geq 0$ et $w_2 = \Phi^{-1}\{(u_2+1)/2\} \geq 0$. Alors, $\phi_r(w_1, w_2) \geq \phi_r(-w_1, w_2)$ pour tout $r \in [0, 1]$. On peut alors écrire

$$\int_0^\rho \{\phi_r(w_1, w_2) - \phi_r(-w_1, w_2)\}\mathrm{d}r \leq \int_0^{\rho'} \{\phi_r(w_1, w_2) - \phi_r(-w_1, w_2)\}\mathrm{d}r.$$

On peut alors conclure que $C^{\chi}_\rho(u_1, u_2) \leq C^{\chi}_{\rho'}(u_1, u_2)$ pour tout $(u_1, u_2) \in [0, 1]^2$.

B.4 Preuve de la Proposition 2.2

Soit Z suivant une loi Normale centrée réduite, c'est-à-dire $Z \sim N(0,1)$. En utilisant les définitions présentées ultérieurement pour $h_a(u)$ et $\tilde{h}_a(u)$, nous obtenons la série d'égalités suivante :

$$\begin{aligned}
\tilde{h}_a(u) - \tilde{h}_a(-u) &= \Phi\{h_a(u)\} - \Phi\{h_a(-u)\} \\
&= \mathbb{P}\{h_a(-u) \leq Z \leq h_a(u)\} \\
&= \mathbb{P}\left\{-\sqrt{G_a^{-1}(u)} - a \leq Z \leq \sqrt{G_a^{-1}(u)} - a\right\} \\
&= \mathbb{P}\left\{(Z+a)^2 \leq G_a^{-1}(u)\right\} \\
&= u.
\end{aligned}$$

B.5 Preuve de la Proposition 2.3

Soient $(X_1, X_2) = ((Z_1 + a_1)^2, (Z_2 + a_2)^2)$ et $(\widetilde{X}_1, \widetilde{X}_2) = ((\widetilde{Z}_1 + a_1)^2, (\widetilde{Z}_2 + a_2)^2)$ des paires de variables aléatoires indépendantes où $(Z_1, Z_2) \sim \Phi_\rho$ et $(\widetilde{Z}_1, \widetilde{Z}_2) \sim \Phi_{\tilde{\rho}}$. De la définition de l'opérateur de concordance, nous avons

$$\begin{aligned}
Q\left(C^{\chi}_{\rho,a_1,a_2}, C^{\chi}_{\tilde{\rho},a_1,a_2}\right) &= 2\mathbb{P}\left\{\left(X_1 - \widetilde{X}_1\right)\left(X_2 - \widetilde{X}_2\right) > 0\right\} - 1 \\
&= 2\mathbb{P}\left\{\left((Z_1 + a_1)^2 - (\widetilde{Z}_1 + a_1)^2\right)\left((Z_2 + a_2)^2 - (\widetilde{Z}_2 + a_2)^2\right) > 0\right\} - 1.
\end{aligned}$$

Il est facile de vérifier que nous avons l'équivalence

$$(Z_j + a_j)^2 - \left(\widetilde{Z}_j + a_j\right)^2 = \left(Z_j - \widetilde{Z}_j\right)\left(Z_j + \widetilde{Z}_j + 2a_j\right).$$

À partir de cette dernière, on peut réécrire l'expression de notre opérateur de concordance sous la forme

$$Q\left(C^{\chi}_{\rho,a_1,a_2}, C^{\chi}_{\tilde{\rho},a_1,a_2}\right) = 2\mathbb{P}(Y_1 Y_2 W_1 W_2 > 0) - 1,$$

où $Y_1 = (Z_1 - \widetilde{Z}_1)/\sqrt{2}, Y_2 = (Z_2 - \widetilde{Z}_2)/\sqrt{2}, W_1 = (Z_1 + \widetilde{Z}_1 + 2a_1)/\sqrt{2}$ et $W_2 = (Z_2 + \widetilde{Z}_2 + 2a_2)/\sqrt{2}$. Il est important de remarquer que $(Y_1, Y_2, W_1, W_2) \sim \Phi^4_\Sigma$. De plus, soit

$$\Omega = \mathbb{I}(Y_1 \leq 0) + \mathbb{I}(Y_2 \leq 0) + \mathbb{I}(W_1 \leq 0) + \mathbb{I}(W_2 \leq 0).$$

Alors, l'événement $Y_1Y_2W_1W_2 > 0$ se produit si et seulement si $\Omega = 0$, $\Omega = 2$ ou $\Omega = 4$. En développant les huit cas possibles pour obtenir une telle situation, nous trouvons l'expression

$$\begin{aligned}
\mathbb{P}(Y_1Y_2W_1W_2 > 0) = \; & 8\mathbb{P}(Y_1 < 0, Y_2 < 0, W_1 < 0, W_2 < 0) \\
& - 4\mathbb{P}(Y_1 < 0, Y_2 < 0, W_1 < 0) - 4\mathbb{P}(Y_1 < 0, Y_2 < 0, W_2 < 0) \\
& - 4\mathbb{P}(Y_1 < 0, W_1 < 0, W_2 < 0) - 4\mathbb{P}(Y_2 < 0, W_1 < 0, W_2 < 0) \\
& + 2\mathbb{P}(Y_1 < 0, Y_2 < 0) + 2\mathbb{P}(W_1 < 0, W_2 < 0) \\
& + 2\mathbb{P}(Y_1 < 0, W_2 < 0) + 2\mathbb{P}(Y_2 < 0, W_1 < 0) \\
& + 2\mathbb{P}(Y_1 < 0, W_1 < 0) + 2\mathbb{P}(Y_2 < 0, W_2 < 0) \\
& + 1 - \mathbb{P}(Y_1 < 0) - \mathbb{P}(Y_2 < 0) - \mathbb{P}(W_1 < 0) - \mathbb{P}(W_2 < 0).
\end{aligned}$$

Le reste de la preuve se fait relativement facilement en effectuant de simples calculs et en utilisant les notations introduites.

B.6 Preuve du Corollaire 2.1

Premièrement, on sait que le vecteur aléatoire $(Z_1^*, Z_2^*, Z_3^*) = (-Z_1, -Z_2, -Z_3)$ possède la même distribution que le vecteur aléatoire $(Z_1, Z_2, Z_3) \sim \Phi^3_{\Sigma'}$. À partir de cette

observation, on obtient

$$\begin{aligned}\Phi^3_{\Sigma'}(0,0,0) &= \mathbb{P}(Z_1 \leq 0, Z_2 \leq 0, Z_3 \leq 0) \\ &= \mathbb{P}(Z_1 > 0, Z_2 > 0) + \mathbb{P}(Z_2 > 0, Z_3 > 0) - \mathbb{P}(Z_1 > 0, Z_2 > 0, Z_3 > 0) - \frac{1}{4} \\ &= \mathbb{P}(Z_1^* < 0, Z_2^* < 0) + \mathbb{P}(Z_2^* < 0, Z_3^* < 0) - \mathbb{P}(Z_1^* < 0, Z_2^* < 0, Z_3^* < 0) - \frac{1}{4} \\ &= \Phi_{\rho^+}(0,0) + \Phi_{\rho^-}(0,0) - \Phi^3_{\Sigma'}(0,0,0) - \frac{1}{4}.\end{aligned}$$

On peut réécrire cette expression sous la forme $\Phi^3_{\Sigma'}(0,0,0) = \frac{1}{2}\Phi_{\rho^+}(0,0) + \frac{1}{2}\Phi_{\rho^-}(0,0) - \frac{1}{8}$. De la formule générale pour l'opérateur de concordance de la copule Khi-deux bivariée, nous obtenons la formule adaptée

$$\begin{aligned}Q\left(C^\chi_\rho, C^\chi_{\rho'}\right) &= Q\left(C^\chi_{\rho,0,0}, C^\chi_{\rho',0,0}\right) \\ &= 16\Phi^4_\Sigma(0,0,0,0) - 32\Phi^3_{\Sigma'}(0,0,0) + 8\Phi_{\rho^+}(0,0) + 8\Phi_{\rho^-}(0,0) - 1.\end{aligned}$$

Le résultat souhaité est obtenu en remplaçant $\Phi^3_{\Sigma'}(0,0,0)$ dans la dernière équation par l'expression trouvée précédemment.

B.7 Preuve de la Proposition 2.4

Soient (X_1, Y_1) et (X_2, Y_2) des paires aléatoires indépendantes issues de la distribution Normale standard ayant comme corrélation ρ. Posons $Z_1 = X_1 - X_2$, $Z_2 = X_1 + X_2 + 2a$, $W_1 = Y_1 - Y_2$ et $W_2 = Y_1 + Y_2 + 2a$. Selon ces nouvelles notations, nous pouvons réécrire le tau de Kendall de la copule Normale comme étant $\tau\left(C^N_\rho\right) = 2\mathbb{P}(Z_1 W_1 > 0) - 1$.

De la définition du tau de Kendall, on trouve

$$\begin{aligned}
\tau\left(C^{\chi}_{\rho,a,a}\right) &= 2\mathbb{P}\left[\left\{(X_1+a)^2-(X_2+a)^2\right\}\left\{(Y_1+a)^2-(Y_2+a)^2\right\}>0\right]-1\\
&= 2\mathbb{P}\left\{(X_1-X_2)(X_1+X_2+2a)(Y_1-Y_2)(Y_1+Y_2+2a)>0\right\}-1\\
&= 2\mathbb{P}(Z_1W_1Z_2W_2>0)-1\\
&= 2\mathbb{P}(Z_1W_1>0,Z_2W_2>0)+2\mathbb{P}(Z_1W_1<0,Z_2W_2<0)-1.
\end{aligned}$$

Comme Z_1 et W_1 sont indépendantes de (Z_2, W_2), le couple (Z_1, W_1) est aussi indépendant de (Z_2, W_2). On obtient donc

$$\begin{aligned}
\tau\left(C^{\chi}_{\rho,a}\right) &= 2\mathbb{P}(Z_1W_1>0,Z_2W_2>0)+2\mathbb{P}(Z_1W_1<0,Z_2W_2<0)-1\\
&= \left\{\tau\left(C^{N}_{\rho}\right)+1\right\}\mathbb{P}(Z_2W_2>0)+\left\{1-\tau\left(C^{N}_{\rho}\right)\right\}\mathbb{P}(Z_2W_2<0)-1\\
&= \tau\left(C^{N}_{\rho}\right)\left\{2\mathbb{P}(Z_2W_2>0)-1\right\}.
\end{aligned}$$

Il reste alors à trouver l'expression explicite de $\mathbb{P}(Z_2W_2 > 0)$. Sachant que

$$(Z_2', W_2') = \left(\sqrt{2}a-\frac{Z_2}{\sqrt{2}},\sqrt{2}a-\frac{W_2}{\sqrt{2}}\right)\sim\Phi_\rho,$$

on trouve

$$\begin{aligned}
\mathbb{P}(Z_2W_2>0) &= \mathbb{P}(Z_2>0,W_2>0)+\mathbb{P}(Z_2<0,W_2<0)\\
&= \mathbb{P}\left(Z_2'<\sqrt{2}a,W_2'<\sqrt{2}a\right)+\mathbb{P}\left(Z_2'>\sqrt{2}a,W_2'>\sqrt{2}a\right)\\
&= 2\mathbb{P}\left(Z_2'<\sqrt{2}a,W_2'<\sqrt{2}a\right)+1-\mathbb{P}\left(Z_2'<\sqrt{2}a\right)-\mathbb{P}\left(W_2'<\sqrt{2}a\right)\\
&= 2\Phi_\rho\left(\sqrt{2}a,\sqrt{2}a\right)+1-2\Phi\left(\sqrt{2}a\right).
\end{aligned}$$

À partir de cette dernière égalité, on obtient

$$\tau\left(C^{\chi}_{\rho,a}\right)=\tau\left(C^{N}_{\rho}\right)\left\{4\Phi_\rho\left(\sqrt{2}a,\sqrt{2}a\right)-4\Phi\left(\sqrt{2}a\right)+1\right\}.$$

B.8 Preuve du Corollaire 2.2

Utilisant le fait que $\tau\left(C_\rho^N\right) = 4\Phi_\rho(0,0) - 1 = (2/\pi)\sin^{-1}(\rho)$ et sachant que $\Phi(0) = 1/2$, on trouve

$$\begin{aligned}\tau\left(C_{\rho,0}^\chi\right) &= \tau\left(C_\rho^N\right)\left\{4\Phi_\rho(0,0) - 4\Phi(0) + 1\right\} \\ &= \tau\left(C_\rho^N\right)\cdot\tau\left(C_\rho^N\right) \\ &= \left(\frac{2}{\pi}\sin^{-1}(\rho)\right)^2.\end{aligned}$$

B.9 Preuve de la Proposition 2.5

Soient ρ_{12}, ρ_{13} et ρ_{23}, les éléments hors de la diagonale de la matrice de corrélation Σ. Les valeurs possibles que peuvent prendre ces éléments sont celles faisant en sorte que la matrice de corrélation Σ soit définie positive et ce, par définition de matrice de corrélation. Alors, ceci s'exprime par $(\rho_{12}, \rho_{13}, \rho_{23;1}) \in (-1,1)^3$, où

$$\rho_{23;1} = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{\left(1-\rho_{12}^2\right)\left(1-\rho_{13}^2\right)}},$$

est le coefficient de corrélation partiel des deuxième et troisième composantes en tenant compte de la première. En effectuant quelques manipulations algébriques sur cette dernière équation, on trouve que la matrice de corrélation est définie positive si et seulement si $\rho_{12}, \rho_{13} \in (-1,1)$ et

$$\rho_{23} \in \left[\rho_{12}\rho_{13} - \sqrt{\left(1-\rho_{12}^2\right)\left(1-\rho_{13}^2\right)}, \rho_{12}\rho_{13} + \sqrt{\left(1-\rho_{12}^2\right)\left(1-\rho_{13}^2\right)}\right].$$

Pour la copule Khi-deux bivariée avec corrélation $\rho_{jj'}$ et paramètre de décentralisation (a,a), on sait que la relation $\rho_{jj'} = g_a^{-1}(\tau_{jj'})$ est valable avec g_a suivant la définition

de l'énoncé du théorème. Donc, nous avons

$$\rho_{12}\rho_{13} - \sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)} \leq \rho_{23} \leq \rho_{12}\rho_{13} + \sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}$$

En utilisant les notations respectives pour γ_1 et γ_2 utilisées dans l'énoncé du théorème ainsi que la dernière relation mentionnée, on trouve une expression simplifiée pour notre inéquation qui est

$$\gamma_1 - \sqrt{\gamma_2} \leq \rho_{23} \leq \gamma_1 + \sqrt{\gamma_2}$$

En appliquant la fonction g_a à chacun des membres de l'inéquation, on a

$$g_a\left(\gamma_1 - \sqrt{\gamma_2}\right) \leq g_a\left(\rho_{23}\right) \leq g_a\left(\gamma_1 + \sqrt{\gamma_2}\right).$$

Du fait que $\rho_{jj'} = g_a^{-1}(\tau_{jj'}) \Rightarrow \tau_{jj'} = g_a(\rho_{jj'})$, on obtient finalement

$$g_a\left(\gamma_1 - \sqrt{\gamma_2}\right) \leq \tau_{23} \leq g_a\left(\gamma_1 + \sqrt{\gamma_2}\right).$$

B.10 Preuve du Corollaire 2.3

En reprenant la définition de la Proposition 2.5 pour la fonction $g_a(x)$, il est possible de la modifier pour l'adapter au cas de la copule Khi-deux centrée où $a = 0$. On obtient donc

$$g_0(x) = \left\{\frac{2}{\pi}\sin^{-1}(x)\right\}^2 \quad \text{et} \quad g_0^{-1}(t) = \sin\left(\frac{\pi}{2}\sqrt{t}\right).$$

À partir de ces expressions, il est facile d'établir les formules pour γ_1 et γ_2 selon leur définition respective énoncée précédemment en travaillant sur les fonctions trigonométriques. On trouve alors

$$\gamma_1 = \sin\left(\frac{\pi}{2}\sqrt{\tau_{12}}\right)\sin\left(\frac{\pi}{2}\sqrt{\tau_{13}}\right) \quad \text{et} \quad \gamma_2 = \cos^2\left(\frac{\pi}{2}\sqrt{\tau_{12}}\right)\cos^2\left(\frac{\pi}{2}\sqrt{\tau_{13}}\right).$$

En faisant usage de l'identité trigonométrique selon laquelle $\sin a \sin b + \cos a \cos b = \cos(a - b)$, on en déduit

$$\gamma_1 - \sqrt{\gamma_2} = -\cos\left\{\frac{\pi}{2}\left(\sqrt{\tau_{12}} + \sqrt{\tau_{13}}\right)\right\} \quad \text{et} \quad \gamma_1 + \sqrt{\gamma_2} = \cos\left\{\frac{\pi}{2}\left(\sqrt{\tau_{12}} - \sqrt{\tau_{13}}\right)\right\}.$$

Une autre identité trigonométrique nous permet d'affirmer que

$$\sin^{-1}\left\{-\cos\left(\frac{\pi}{2}r\right)\right\} = -\frac{\pi}{2}(1 - |r|) \quad \text{et} \quad \sin^{-1}\left\{\cos\left(\frac{\pi}{2}r\right)\right\} = \frac{\pi}{2}(1 - |r|),$$

et ce, pour tout $r \in \mathbb{R}$. Ceci nous permet d'obtenir

$$g_0\left(\gamma_1 - \sqrt{\gamma_2}\right) = \left\{\max\left(0, \sqrt{\tau_{12}} + \sqrt{\tau_{13}} - 1\right)\right\}^2 \quad \text{et} \quad g_0\left(\gamma_1 + \sqrt{\gamma_2}\right) = \left(1 - \left|\sqrt{\tau_{12}} - \sqrt{\tau_{13}}\right|\right)^2.$$

Finalement, en se servant du résultat de la Proposition 2.5, on a

$$\begin{aligned} g_a\left(\gamma_1 - \sqrt{\gamma_2}\right) \leq \tau_{23} &\leq g_a\left(\gamma_1 + \sqrt{\gamma_2}\right) \\ g_0\left(\gamma_1 - \sqrt{\gamma_2}\right) \leq \tau_{23} &\leq g_0\left(\gamma_1 + \sqrt{\gamma_2}\right) \\ \left\{\max\left(0, \sqrt{\tau_{12}} + \sqrt{\tau_{13}} - 1\right)\right\}^2 \leq \tau_{23} &\leq \left(1 - \left|\sqrt{\tau_{12}} - \sqrt{\tau_{13}}\right|\right)^2. \end{aligned}$$